PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification [4] :
C12P 21/00, C07H 17/00
C07K 15/04, A61K 39/395

A1

(11) International Publication Number: WO 88/ 01649

(43) International Publication Date: 10 March 1988 (10.03.88)

(21) International Application Number: PCT/US87/02208

(22) International Filing Date: 2 September 1987 (02.09.87)

(31) Priority Application Number: 902,971

(32) Priority Date: 2 September 1986 (02.09.86)

(33) Priority Country: US

(71) Applicant: GENEX CORPORATION [US/US]; 16020 Industrial Drive, Gaithersburg, MD 20877 (US).

(72) Inventors: LADNER, Robert, Charles ; 3827 Green Valley, Iamsville, MD 21754 (US). BIRD, Robert, Earl ; 3903 Morrell Court, Kensington, MD 20895 (US).

(74) Agents: GOLDSTEIN, Jorge, A. et al.; Saidman, Sterne, Kessler & Goldstein, 1225 Connecticut Avenue, N.W., Washington, DC 20036 (US).

(81) Designated States: AT (European patent), BE (European patent), CH (European patent), DE (European patent), DK, FI, FR (European patent), GB (European patent), IT (European patent), JP, LU (European patent), NL (European patent), NO, SE (European patent).

Published
With international search report.

(54) Title: SINGLE POLYPEPTIDE CHAIN BINDING MOLECULES

(57) Abstract

A single polypeptide chain binding molecule which has binding specificity substantially similar to the binding specificity of the light and heavy chain aggregate variable region of an antibody, genetic sequences coding therefor, recombinant DNA methods of producing such molecule and uses for such molecule.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	ML	Mali
AU	Australia	GA	Gabon	MR	Mauritania
BB	Barbados	GB	United Kingdom	MW	Malawi
BE	Belgium	HU	Hungary	NL	Netherlands
BG	Bulgaria	IT	Italy	NO	Norway
BJ	Benin	JP	Japan	RO	Romania
BR	Brazil	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	LI	Liechtenstein	SN	Senegal
CH	Switzerland	LK	Sri Lanka	SU	Soviet Union
CM	Cameroon	LU	Luxembourg	TD	Chad
DE	Germany, Federal Republic of	MC	Monaco	TG	Togo
DK	Denmark	MG	Madagascar	US	United States of America
FI	Finland				

SINGLE POLYPEPTIDE CHAIN BINDING MOLECULES

BACKGROUND OF THE INVENTION

This application is a continuation-in-part of Application Serial No. 902,971, filed September 2, 1986, the contents of which are herein fully incorporated by reference.

Field of the Invention

The present invention relates to single polypeptide chain binding molecules having the three dimensional folding, and thus the binding ability and specificity, of the variable region of an antibody. Methods of producing these molecules by genetic engineering are also disclosed.

Description of the Background Art

The advent of modern molecular biology and immunology has brought about the possibility of producing large quantities of biologically active materials in highly reproduceable form and with low cost. Briefly, the gene sequence coding for a desired natural protein is isolated, replicated (cloned) and introduced into a foreign host such as a bacterium, a yeast (or other fungi) or a mammalian cell line in culture, with appropriate regulatory control signals. When the signals are activated, the gene is transcribed and translated, and expresses the desired protein. In this manner, such useful biologically active materials as hormones, enzymes or antibodies have been cloned and expressed in foreign hosts.

One of the problems with this approach is that it is limited by the "one gene, one polypeptide chain" principle of molecular biology. In other words, a genetic sequence codes for a single polypeptide chain. Many biologically active polypeptides, however, are aggregates of two or more chains. For example, antibodies are three-dimensional aggregates of two heavy and two light chains. In the same manner, large enzymes such as aspartate transcarbamylase, for example, are aggregates of six catalytic and six regulatory chains, these chains being different. In order to produce such complex materials by recombinant DNA technology in foreign hosts, it becomes necessary to clone and express a gene coding for each one of the different kinds of polypeptide chains. These genes can be expressed in separate hosts. The resulting polypeptide chains from each host would then have to be reaggregated and allowed to refold together in solution. Alternatively, the two or more genes coding for the two or more polypeptide chains of the aggregate could be expressed in the same host simultaneously, so that refolding and reassociation into the native structure with biological activity will occur after expression. The approach, however, necessitates expression of multiple genes, and as indicated, in some cases, in multiple and different hosts. These approaches have proved to be inefficient.

Even if the two or more genes are expressed in the same organism it is quite difficult to get them all expressed in the required amounts.

A classical example of multigene expression to form multimeric polypeptides is the expression by recombinant DNA technology of antibodies. Genes for heavy and light chains have been introduced into ap-

propriate hosts and expressed, followed by reaggregation of these individual chains into functional antibody molecules (see for example Munro, Nature, 312:597 (1984); Morrison, S.L. Science 229:1202 (1985); Oi et al., BioTechniques 4:214 (1986)); Wood et al., Nature, 314: 446-449 (1985)).

Antibody molecules have two generally recognized regions, in each of the heavy and light chains. These regions are the so-called "variable" region which is responsible for binding to the specific antigen in question, and the so-called "constant" region which is responsible for biological effector responses such as complement binding, etc. The constant regions are not necessary for antigen binding. The constant regions have been separated from the antibody molecule, and biologically active (i.e. binding) variable regions have been obtained.

The variable regions of an antibody are composed of a light chain and a heavy chain. Light and heavy chain variable regions have been cloned and expressed in foreign hosts, and maintain their binding ability (Moore et al, European Patent Publication 0088994 (published September 21, 1983)).

Further, it is by now well established that all antibodies of a certain class and their Fab fragments whose structures have been determined by X-ray crystallography, even when from different species, show closely similar variable regions despite large differences in the hypervariable segments. The immunoglobulin variable region seems to be tolerant toward mutations in the combining loops. Therefore, other than in the hypervariable regions, most of the so

called "variable" regions of antibodies, which are defined by both heavy and light chains, are in fact quite constant in their three dimensional arrangement. See, for example, Huber, R., "Structural Basis for Antigen-Antibody Recognition," Science, 233:702-703 (1986).

It would be very efficient if one could produce single polypeptide-chain molecules which have the same biological activity as the multiple chain aggregates such as, for example, multiple chain antibody aggregates or enzyme aggregates. Given the "one gene-one-polypeptide chain" principle, such single chain molecules would be more readily produceable, and would not necessitate multiple hosts or multiple genes in the cloning and expression. In order to accomplish this, it is first necessary to devise a method for generating single chain structures from two-chain aggregate structures, wherein the single chain will retain the three-dimensional folding of the separate natural aggregate of two polypeptide chains.

While the art has discussed the study of proteins in three dimensions, and has suggested modifying their architecture (see, for example, the article "Protein Architecture: Designing from the Ground Up," by Van Brunt, J., BioTechnology, 4: 277-283 (April, 1986)), the problem of generating single chain structures from multiple chain structures, wherein the single chain structure will retain the three-dimensional architecture of the multiple chain aggregate, has not been satisfactorily addressed.

Given that methods for the preparation of genetic sequences, their replication, their linking to expres-

sion control regions, formation of vectors therewith and transformation of appropriate hosts are well understood techniques, it would indeed be greatly advantageous to be able to produce, by genetic engineering, single polypeptide chain binding proteins having the characteristics and binding ability of multi chain variable regions of antibody molecules.

SUMMARY OF THE INVENTION

The present invention starts with a computer based system and method to determine chemical structures for converting two naturally aggregated but chemically separated light and heavy polypeptide chains from an antibody variable region into a single polypeptide chain which will fold into a three dimensional structure very similar to the original structure made of the two polypeptide chains.

The single polypeptide chain obtained from this method can then be used to prepare a genetic sequence coding therefor. The genetic sequence can then be replicated in appropriate hosts, further linked to control regions, and transformed into expression hosts, wherein it can be expressed. The resulting single polypeptide chain binding protein, upon refolding, has the binding characteristics of the aggregate of the original two (heavy and light) polypeptide chains of the variable region of the antibody.

The invention therefore comprises:

A single polypeptide chain binding molecule which has binding specificity substantially similar to the binding specificity of the light and heavy chain aggregate variable region of an antibody.

The invention also comprises genetic sequences coding for the above mentioned single polypeptide chain, cloning and expression vectors containing such genetic sequences, hosts transformed with such vectors, and methods of production of such polypeptides by expression of the underlying genetic sequences in such hosts.

The invention also extends to uses for the binding proteins, including uses in diagnostics, therapy, in vivo and in vitro imaging, purifications, and biosensors. The invention also extends to the single chain binding molecules in immobilized form, or in detectably labelled forms for utilization in the above mentioned diagnostic, imaging, purification or biosensor applications. It also extends to conjugates of the single polypeptide chain binding molecules with therapeutic agents such as drugs or specific toxins, for delivery to a specific site in an animal, such as a human patient.

Essentially all of the uses that the prior art has envisioned for monoclonal or polyclonal antibodies, or for variable region fragments thereof, can be considered for the molecules of the present invention.

The advantages of single chain over conventional antibodies are smaller size, greater stability and significantly reduced cost. The smaller size of single chain antibodies may reduce the body's immunologic reaction and thus increase the safety and efficacy of therapeutic applications. Conversely, the single chain antibodies could be engineered to be highly antigenic. The increased stability and lower cost permits greater use in biosensors and protein purifica-

tion systems. Because it is a smaller and simpler protein, the single chain antibody is easier to further modify by protein engineering so as to improve both its binding affinity and its specificity. Improved affinity will increase the sensitivity of diagnosis and detection and detection systems while improved specificity will reduce the number of false positives observed.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention as defined in the claims can be better understood with reference to the text and to the following drawings, as follows:

Figure 1 is a block diagram of the hardware aspects of the serial processor mode of the present invention.

Figure 2 is a block diagram of an alternate embodiment of the hardware aspects of the present invention.

Figure 3 is a block diagram of the three general steps of the present invention.

Figure 4 is a block diagram of the steps in the site selection step in the single linker embodiment.

Figure 5A is a schematic two dimensional simplified representation of the light chain L and heavy chain H of two naturally aggregated antibody variable region F_v polypeptide chains used to illustrate the site selection process.

Figure 5B is a two dimensional representation of the three dimensional relationship of the two aggregated polypeptide chains showing the light chain L (----) and the heavy chain H (-) of the variable region of one antibody.

Figure 6A is a simplified two dimensional schematic diagram of the two polypeptide chains showing the location of the residue Tau 1 and the residue Sigma 1.

Figure 6B is a two dimensional representation of the actual relationship of the two polypeptide chains showing the residue Tau 1 and the residue Sigma 1.

Figure 7 shows in very simplified schematic way the concept of the direction linkers that are possible between the various possible sites on the light chain L and the heavy chain H in the residue Tau 1 and residue Sigma 1 respectively.

Figure 8A is a two dimensional simplified schematic diagram of a single chain antibody linking together two separate chains ((Heavy ———) and (light _ _ _)) by linker 1 (----) to produce a single chain antibody.

Figure 8B is a two dimensional representation showing a single chain antibody produced by linking two aggregated polypeptide chains using linker 1.

Figure 9 shows a block diagram of candidate selection for correct span.

Figure 10 shows a block diagram of candidate selection for correct direction from N terminal to C terminal.

Figure 11 shows a comparison of direction of a gap to direction of a candidate.

Figure 12 shows a block diagram of candidate selection for correct orientation at both ends.

Figure 13 shows a block diagram of selection of sites for the two-linker embodiment.

Figure 14 shows examples of rules by which candidates may be ranked.

Figure 15A shows a two-dimensional simplified representation of the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the first two sites to be linked.

Figure 15B shows a two-dimensional representation of the three-dimensional relationships between the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the regions in which the second sites to be linked can be found and the linker between the first pair of sites.

Figure 16A shows the two-dimensional simplified representation of the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the regions in which the second sites to be linked can be found and the linker between the first pair of sites.

Figure 16B shows the two-dimensional representation of the three-dimensional relationships between the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the regions in which the second sites to be linked can be found and the linker between the first pair of sites.

Figure 17A shows the two-dimensional simplified representation of the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the second linker and the portions of the native protein which are lost.

Figure 17B shows the two-dimensional representation of the three-dimensional relationships between the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the second linker and the portions of native protein which are lost.

Figure 18 shows the two-dimensional simplified representation of the variable domain of an Fv light chain, L, and the variable domain of an Fv heavy chain, H, showing the complete construction.

Figure 19 shows a block diagram of the parallel processing mode of the present invention.

Figure 20A shows five pieces of molecular structure. The uppermost segment consists of two peptides joined by a long line. The separation between the peptides is 12.7 A. The first C of each peptide lies on the X-axis. The two dots indicate the standard reference point in each peptide.

Below the gap are four linker candidates (labeled 1,2,3 & 4), represented by a line joining the alpha carbons. In all cases, the first and penultimate alpha carbons are on lines parallel to the X-axis, spaced 8.0 A apart. Note that the space between dots in linker 1 is much shorter than in the gap.

Figure 20B shows the initial peptides of linkers 2, 3, and 4 which have been aligned with the first peptide of the gap. For clarity, the linkers have been translated vertically to their original positions.

The vector from the first peptide in the gap to the second peptide in the gap lies along the X-axis, a corresponding vector for linkers 3 and 4 also lies along the X-axis. Linker 2, however, has this vector pointing up and to the right, thus linker 2 is rejected.

Figure 20C shows the ten atoms which compose the initial and final peptides of linkers 3 and 4, which have been least-squares fit to the corresponding atoms from the gap. These peptides have been drawn in.

Note that in the gap and in linker 4 the final peptide points down and lies more-or-less in the plane of the paper. In linker 3, however, this final peptide points down and to the left and is twisted about 90 degrees so that the carbonyl oxygen points toward the viewer. Thus linker 3 is rejected.

Sections B and C are stereo diagrams which may be viewed with the standard stereo viewer provided.

Figure 21 shows the nucleotide sequence and translation of the sequence for the heavy chain of a mouse anti bovine growth hormone (BGH) monoclonal antibody.

Figure 22 shows the nucleotide sequence and translation of the sequence for the light chain of the same monoclonal antibody as that shown in Figure 21.

Figure 23 is a plasmid restriction map containing the variable heavy chain sequence (pGX3772) and that containing the variable light sequence (pGX3773) shown in figures 21 and 22.

Figure 24 shows construction TRY40 comprising the nucleotide sequence and its translation sequence of a single polypeptide chain binding protein prepared according to the methods of the invention.

Figure 25 shows a restriction map of the expression vector pGX3776 carrying a single chain binding protein, the sequence of which is shown in Figure 24. In this and subsequent plasmid maps (Figures 27 and 29) the hashed bar represents the promoter O_L/P_R sequence and the solid bar represents heavy chain variable region sequences.

Figure 26 shows the sequences of TRY61, another single chain binding protein of the invention.

Figure 27 sh ws expression plasmid pGX4904 carrying the genetic sequ nce sh wn in Figure 26.

Figure 28 shows the sequences of TRY59, another single chain binding protein of the invention.

Figure 29 shows the expression plasmid pGX <u>4908</u> carrying the genetic sequence shown in Figure 28.

Figures 30A, 30B, 30C, and 30D (stereo) are explained in detail in Example 1. They show the design and construction of double linked single chain antibody TRY40.

Figures 31A and 31B (stereo) are explained in detail in Example 2. They show the design and construction of single linked single chain antibody TRY61.

Figures 32A and 32B (stereo) are explained in detail in Example 3. They show the design and construction of single linked single chain antibody TRY59.

Figure 33 is explained in Example 4 and shows the sequence of TRYι04b.

Figure 34 shows a restriction map of the expression vector pGX4910 carrying a single linker construction, the sequence of which is shown in Figure 33.

Figure 35 shows the assay results for BGH binding activity wherein strip one represents TRY61 and strip two represents TRY40.

Figure 36 is explained in Example 4 and shows the results of competing the F_{ab} portion of 3C2 monoclonal with TRY59 protein.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

TABLE OF CONTENTS

I. General Overview

The present invention starts with a computer based system and method for determining and displaying possible chemical structures (linkers) for converting two naturally aggregated but chemically separate heavy and light (H and L) polypeptide chains from the variable region of a given antibody into a single polypeptide chain which will fold into a three dimensional struc-

ture very similar to the original structure made of two polypeptide chains. The original structure is referred to hereafter as "native protein."

The first general step of the three general design steps of the present invention involves selection of plausible sites to be linked. In the case of a single linker, criteria are utilized to select a plausible site on each of the two polypeptide chains (H and L in the variable region) which will result in 1) a minimum loss of residues from the native protein chains and 2) a linker of minimum number of amino acids consistent with the need for stability. A pair of sites defines a gap to be bridged or linked.

A two-or-more-linker approach is adopted when a single linker can not achieve the two stated goals. In both the single-linker case and the two-or-more-linker case, more than one gap may be selected for use in the second general step.

The second general step of the present invention involves examining a data base to determine possible linkers to fill the plausible gaps selected in the first general step, so that candidates can be enrolled for the third general step. Specifically, a data base contains a large number of amino acid sequences for which the three-dimensional structure is known. In the second general step, this data base is examined to find which amino acid sequences can bridge the gap or gaps to create a plausible one-polypeptide structure which retains most of the three dimensional features of the native (i.e. original aggregate) variable region molecule. The testing of each possible linker proceeds in three general substeps. The first general substep utilizes the length of the possible candidate.

Specifically, the span r length (a scalar quantity) of the candidate is compared to the span of each of the gaps. If the difference between the length of the candidate and the span of any one of the gaps is less than a selected quantity, then the present invention proceeds to the second general substep with respect to this candidate. Figure 20A shows one gap and four possible linkers. The first linker fails the first general substep because its span is quite different from the span of the gap.

In the second general substep, called the direction substep, the initial peptide of the candidate is aligned with the initial peptide of each gap. Specifically, a selected number of atoms in the initial peptide of the candidate are rotated and translated as a rigid body to best fit the corresponding atoms in the initial peptide of each gap. The three dimensional vector (called the direction of the linker) from the initial peptide of the candidate linker to the final peptide of the candidate linker is compared to the three dimensional vector (call the direction of the gap) from the initial peptide of each gap to the final peptide of the same gap. If the ends of these two vectors come within a preselected distance of each other, the present invention proceeds to the third general substep of the second general step with respect to this candidate linker.

Figure 20B shows one gap and three linkers. All the linkers have the correct span and the initial peptides have been aligned. The second linker fails the second general substep because its direction is quite different from that of the gap; the other two linkers are carried forward to the third general substep of the second general step.

In the third general substep of the second design of the step of the present invention, the orientations of the terminal peptides of each linker are compared to the orientations of the terminal peptides of each gap. Specifically, a selected number of atoms (3, 4, or 5, 5 in the prefered embodiment) from the initial peptide of the candidate plus the same selected number of atoms (3, 4, or 5; 5 in the prefered embodiment) from the final peptide of the candidate are taken as a rigid body. The corresponding atoms from one of the gaps (viz 5 from the initial peptide and 5 from the final peptide) are taken as a second rigid body. These two rigid bodies are superimposed by a least-squares fit. If the error for this fit is below some preselected value, then the candidate passes the third general substep of the second general step and is enrolled for the third general step of the present invention. If the error is greater than or equal to the preselected value, the next gap is tested. When all gaps have been tested without finding a sufficiently good fit, the candidate is abandoned.

The third general step of the present invention results in the ranking of the linker candidates from most plausible to least plausible. The most plausible candidate is the fragment that can bridge the two plausible sites of one of the gaps to form a single polypeptide chain, where the bridge will least distort the resulting three dimensional folding of the single polypeptide chain from the natural folding of the aggregate of the two originally chemically separate chains.

In this third general step of the present invention, an expert operator uses an interactive computer-graphics approach to rank the linker candidates from

most plausible to least plausible. This ranking is done by observing the interactions between the linker candidate with all retained portions of the native protein. A set of rules are used for the ranking. These expert system rules can be built into the system so that the linkers are displayed only after they have satisfied the expert system rules that are utilized.

The present invention can be programmed so that certain expert rules are utilized as a first general substep in the third general step to rank candidates and even eliminate unsuitable candidates before visual inspection by an expert operator, which would be the second general substep of the third general step. These expert rules assist the expert operator in ranking the candidates from most plausible to least plausible. These expert rules can be modified based on experimental data on linkers produced by the system and methods of the present invention.

The most plausible candidate is a genetically producible single polypeptide chain binding molecule which has a very significantly higher probability (a million or more as compared to a random selection) of folding into a three dimensional structure very similar to the original structure made of the heavy and light chains of the antibody variable region than would be produced if random selection of the linker was done. In this way, the computer based system and method of the present invention can be utilized to engineer single polypeptide chains by using one or more linkers which convert naturally aggregated but chemically separated polypeptide chains into the desired single chain.

The elected candidate offers to the user a linked chain structure having a very significantly increased

probability of proper folding than would be obtained using a random selection process. This means that the genetic engineering aspect of creating the desired single polypeptide chain is significantly reduced, since the number of candidates that have to be genetically engineered in practice is reduced by a corresponding amount. The most plausible candidate can be used to genetically engineer an actual molecule.

The parameters of the various candidates can be stored for later use. They can also be provided to the user either visually or recorded on a suitable media (paper, magnetic tape, color slides, etc.). The results of the various steps utilized in the design process can also be stored for later use or examination.

The design steps of the present invention operate on a conventional minicomputer system having storage devices capable of storing the amino acid sequence-structure data base, the various application programs utilized and the parameters of the possible linker candidates that are being evaluated.

The minicomputer CPU is connected by a suitable serial processor structure to an interactive computer-graphics display system. Typically, the interactive computer-graphics display system comprises a display terminal with resident three-dimensional application software and associated input and output devices, such as X/Y plotters, position control devices (potentiometers, an x-y tablet, or a mouse), and keyboard.

The interactive computer-graphics display system allows the expert operator to view the chemical structures being evaluated in the design process of the

present invention. Graphics and programs are used to select the gaps (Gen. Step 1), and to rank candidates (Gen. Step 3). Essentially, it operates in the same fashion for the single linker embodiment and for the two or more linker embodiments.

For example, during the first general step of the present invention, the computer-graphics interactive display system allows the expert operator to visually display the two naturally aggregated but chemically separate polypeptide chains. Using three dimensional software resident in the computer-graphics display system, the visual representation of the two separate polypeptide chains can be manipulated as desired. For example, the portion of the chain(s) being viewed can be magnified electronically, and such magnification can be performed in a zoom mode. Conversely, the image can be reduced in size, and this reduction can also be done in a reverse zoom mode. The position of the portion of the molecule can be translated, and the displayed molecule can be rotated about any one of the three axes (x, y and z). Specific atoms in the chain can be selected with an electronic pointer. Selected atoms can be labeled with appropriate text. Specific portions of native protein or linker can be identified with color or text or brightness. Unwanted portions of the chain can be erased from the image being displayed so as to provide the expert operator with a visual image that represents only a selected aspect of the chain(s). Atoms selected by pointing or by name can be placed at the center of the three dimensional display; subsequent rotation uses the selected atom as the origin. These and other display aspects provide

the expert operator with the ability to visually represent portions of the chains which increase the ability to perform the structural design process.

One of the modes of the present invention utilizes a serial computational architecture. This architecture using the present equipment requires approximately four to six hours of machine and operator time in order to go through the various operations required for the three general steps for a particular selection of gaps. Obviously, it would be desirable to significantly reduce the time since a considerable portion thereof is the time it takes for the computer system to perform the necessary computational steps.

An alternate embodiment of the present invention utilizes a parallel processing architecture. This parallel processing architecture significantly reduces the time required to perform the necessary computational steps. A hypercube of a large number of nodes can be utilized so that the various linkers that are possible for the selected sites can be rapidly presented to the expert system operator for evaluation.

Since there are between 200 and 300 known protein structures, the parallel processing approach can be utilized. There currently are computers commercially available that have as many as 1,024 computing nodes.

Using a parallel processing approach, the data base of observed peptide structures can be divided into as many parts as there are computing nodes. For example, if there are structures for 195 proteins with 219 amino acids each, one would have structures for 195x218 dipeptides, 195x217 tripeptides, 195x216 tetrapeptides, etc. One can extract all peptides up to

some length n. For example, if n were 30, one would have 195x30x204 peptides. Of course, proteins vary in length, but with 100 to 400 proteins of average length 200 (for example), and for peptide linkers up to length 30 amino acids (or any other reasonable number), one will have between 1,000,000 and 4,000,000 peptide structures. Once the peptides have been extracted and labeled with the protein from which they came, one is free to divide all the peptides as evenly as possible among the available computing nodes.

The parallel processing mode operates as follows. The data base of known peptides is divided among the available nodes. Each gap is sent to all the nodes. Each node takes the gap and tests it against those peptides which have been assigned to it and returns information about any peptides which fit the gap and therefore are candidate linkers. As the testing for matches between peptides and gaps proceeds independently in each node, the searching will go faster by a factor equal to the number of nodes.

A first embodiment of the present invention utilizes a single linker to convert the naturally aggregated but chemically separate heavy and light chains into a single polypeptide chain which will fold into a three dimensional structure very similar to the original structure made of two polypeptide chains.

A second embodiment utilizes two or more linkers to convert the two heavy and light chains into the desired single polypeptide chain. The steps involved in each of these embodiments utilizing the present invention are illustrated in the explanation below.

Once the correct amino acid sequence for a single chain binding protein has been defined by the computer assisted methodology, it is possible, by methods well known to those with skill in the art, to prepare an underlying genetic sequence coding therefor.

In preparing this genetic sequence, it is possible to utilize synthetic DNA by synthesizing the entire sequence de novo. Alternatively, it is possible to obtain cDNA sequences coding for certain preserved portions of the light and heavy chains of the desired antibody, and splice them together by means of the necessary sequence coding for the peptide linker, as described.

Also by methods known in the art, the resulting sequence can be amplified by utilizing well known cloning vectors and well known hosts. Furthermore, the amplified sequence, after checking for correctness, can be linked to promoter and terminator signals, inserted into appropriate expression vectors, and transformed into hosts such as procaryotic or eucaryotic hosts. Bacteria, yeasts (or other fungi) or mammalian cells can be utilized. Upon expression, either by itself or as part of fusion polypeptides, as will otherwise be known to those of skill in the art, the single chain binding protein is allowed to refold in physiological solution, at appropriate conditions of pH, ionic strength, temperature, and redox potential, and purified by standard separation procedures. These would include chromatography in its various different types, known to those with skill in the art.

The thus obtained purified single chain binding protein can be utilized by itself, in detectably la-

belled form, in immobilized form, or conjugated to drugs or other appropriate therapeutic agents, in diagnostic, imaging, biosensors, purifications, and therapeutic uses and compositions. Essentially all uses envisioned for antibodies or for variable region fragments thereof can be considered for the molecules of the present invention.

II. Hardware and Software Environment

A block diagram of the hardware aspects of the present invention is found in Figure 1. A central processing unit (CPU) 102 is connected to a first bus (designated massbus 104) and to a second bus (designated Unibus 106). A suitable form for CPU 102 is a model Vax 11/780 made by Digital Equipment Corporation of Maynard, Massachusetts. Any suitable type of CPU, however, can be used.

Bus 104 connects CPU 102 to a plurality of storage devices. In the best mode, these storage devices include a tape drive unit 106. The tape drive unit 106 can be used, for example, to load into the system the data base of the amino acid sequences whose three dimensional structures are known. A suitable form for tape drive 106 is a Digital Equipment Corporation model TU 78 drive, which operates at 125 inches per second, and has a 1600-6250 bit per inch (BPI) dual capability. Any suitable type of tape drive can be used, however.

Another storage device is a pair of hard disk units labeled generally by reference numeral 108. A suitable form for disk drive 108 comprises two Digital Equipment Corporation Rm05 disk drives having, for

example, 256 Mbytes of storage per disk. Another disk drive system is also provided in the serial processor mode and is labeled by reference numeral 110. This disk drive system is also connected to CPU 102 by bus 104. A suitable form for the disk system 110 comprises three Digital Equipment Corporation model Ra 81 hard disk drives having, for example, 450 Mbytes of storage per disk.

Dynamic random access memory is also provided by a memory stage 112 also connected to CPU 102 by bus 104. Any suitable type of dynamic memory storage device can be used. In the serial processor mode, the memory is made up of a plurality of semi- conductor storage devices found in a DEC model Ecc memory unit. Any suitable type of dynamic memory can be employed.

The disk drives 108 and 110 store several different blocks of information. For example, they store the data base containing the amino acid sequences and structures that are read in by the tape drive 106. They also store the application software package required to search the data base in accordance with the procedures of the present invention. They also store the documentation and executables of the software. The hypothetical molecules that are produced and structurally examined by the present invention are represented in the same format used to represent the protein structures in the data base. Using this format, these hypothetical molecules are also stored by the disk drives 108 and 110 for use during the structural design process and for subsequent use after the process has been completed.

A Digital Equipment Corporation VAX/VMS DEC operating system allows for multiple users and assures file system integrity. It provides virtual memory, which relieves the programer of having to worry about the amount of memory that is used. Initial software was developed under versions 3.0 to 3.2 of the VAX/VMS operating system. The serial processor mode currently is running on version 4.4. DEC editors and FORTRAN compiler were utilized.

The CPU 102 is connected by Bus 106 to a multiplexer 114. The multiplexer allows a plurality of devices to be connected to the CPU 102 via Bus 106. A suitable form for multiplexer 114 is a Digital Equipment Corporation model Dz 16 terminal multiplexer. In the preferred embodiment, two of these multiplexers are used. The multiplexer 114 supports terminals (not shown in Figure 1) and the serial communications (at 19.2 Kbaud, for example) to the computer-graphics display system indicated by the dash lined box 116.

The computer-graphics display system 116 includes an electronics stage 118. The electronic stage 118 is used for receiving the visual image prepared by CPU 102 and for displaying it to the user on a display (typically one involving color) 120. The electronic stage 118 in connection with the associated subsystems of the computer-graphics display system 116 provide for local control of specific functions, as described below. A suitable form of the electronics system 118 is a model PS 320 made by Evans & Sutherland Corp. of Salt Lake, Utah. A suitable form for the display 120 is either a 25 inch color monitor or a 19 inch color monitor from Evans & Sutherland.

Dynamic random access memory 122 is connected to the electronic stage 118. Memory 122 allows the electronic system 118 to provide the local control of the image discussed below. In addition, a keyboard 124 of conventional design is connected to the electronic stage 118, as is an x/y tablet 126 and a plurality of dials 128. The keyboard 124, x/y tablet 126, and dials 128 in the serial processor mode are also obtained from Evans & Sutherland.

The computer generated graphics system 116, as discussed above, receives from CPU 102 the image to be displayed. It provides local control over the displayed image so that specific desired user initiated functions can be performed, such as:

(1) zoom (so as to increase or decrease the size of the image being displayed;

(2) clipping (where the sides, front or back of the image being displayed are removed);

(3) intensity depth queing (where objects further away from the viewer are made dimmer so as to provide a desired depth effect in the image being displayed);

(4) translation of the image in any of the three axes of the coordinate system utilized to plot the molecules being displayed;

(5) rotation in any of the three directions of the image being displayed;

(6) on/off control of the logical segments of the picture. For example, a line connecting the alpha carbons of the native protein might be one logical segment; labels on some or all of the residues of the native protein might be a second logical segment; a trace of the alpha carbons of the linker(s) might be a

third segment; and a stick figure connecting Carbon, Nitrogen, Oxygen, and Sulphur atoms of the linker(s) and adjacent residue of the native protein might be a fourth logical segment. The user seldom wants to see all of these at once; rather the operator first becomes oriented by viewing the first two segments at low magnification. Then the labels are switched off and the linker carbon trace is turned on. Once the general features of the linker are seen, the operator zooms to higher magnification and turns on the segments which hold more detail;

(7) selection of atoms in the most detailed logical segment. Despite the power of modern graphics, the operator can be overwhelmed by too much detail at once. Thus the operator will pick one atom and ask to see all amino acids within some radius of that atom, typically 6 Angstroms, but other radii can be used. The user may also specify that certain amino acids will be included in addition to those that fall within the specified radius of the selected atom;

(8) changing of the colors of various portions of the image being displayed so as to indicate to the viewer particular information using visual queing.

As stated above, the serial processor mode of the present invention currently is running the application software on version 4.4 of the Vax/Vms operating system used in conjunction with CPU 102. The application programs were programmed using the FLECS (FORTRAN Language with Extended Control Sections) programming language written in 1974 by Terry Beyer of the University of Oregon, Eugene, Oregon. FLECS is a FORTRAN preprocessor, which allows more logical programming.

All of the code used in the serial processor mode was developed in FLECS. It can be appreciated, however, that the present invention encompasses other operating systems and programming languages.

The macromolecules displayed on color display 120 of the computer-graphics display system 116 utilize an extensively modified version of version 5.6 of FRODO. FRODO is a program for displaying and manipulating macromolecules. FRODO was written by T.A. Jones at Max Planck Institute for Biochemistry, Munich, West Germany, for building or modeling in protein crystallography. FRODO version 5.6 was modified so as to be driven by command files; programs were then written to create the command files. It is utilized by the electronic stage 118 to display and manipulate images on the color display 120. Again, any suitable type of program can be used for displaying and manipulating the macromolecules, the coordinates of which are provided to the computer-graphics display system 116 by the CPU 102.

Design documentation and memos were written using PDL (Program Design Language) from Caine, Farber & Gordon of Pasadena, California. Again, any suitable type of program can be used for the design documents and memos.

Figure 2 shows a block diagram for an improved version of the hardware system of the present invention. Like numbers refer to like items of Figure 1. Only the differences between the serial processor mode system of Figure 1 and the improved system of Figure 2 are discussed below.

The CPU 102' is the latest version of the Vax 11/780 from Digital Equipment Corporation. The latest processor from DEC in the VAX product family is approximately ten times faster than the version shown in the serial processor mode of Figure 1.

Instead of the two Rm05 disk drives 108 of Figure 1, the embodiment of Figure 2 utilizes five RA 81 disk drive units 110'. This is to upgrade the present system to more state of the art disk drive units, which provide greater storage capability and faster access.

Serial processor 106 is connected directly to the electronic stage 118' of the computer-graphics display system 116. The parallel interface in the embodiment of Figure 2 replaces the serial interface approach of the serial processor mode of Figure 1. This allows for faster interaction between CPU 102' and electronic stage 118' so as to provide faster data display to the expert operator.

Disposed in front of color display 120 is a stereo viewer 202. A suitable form for stereo viewer 202 is made by Terabit, Salt Lake City, Utah. Stereo viewer 202 would provide better 3-D perception to the expert operator than can be obtained presently through rotation of the molecule.

In addition, this embodiment replaces the FRODO macromolecule display programs with a program designed to show a series of related hypothetical molecules. This newer program performs the operations more quickly so that the related hypothetical molecules can be presented to the expert operator in a short enough time that makes examination less burdensome on the operator.

The programs can be modified so as to cause the present invention to eliminate candidates in the second general step where obvious rules have been violated by the structures that are produced. For example, one rule could be that if an atom in a linker comes closer than one Angstrom to an atom in the native structure the candidate would be automatically eliminated.

In addition, the surface accessibility of molecules could be determined and a score based on the hydrophobic residues in contact with the solvent could be determined. After the hydrophobic residues have been calculated, the candidates could be ranked so that undesired candidates could automatically be eliminated. The protein is modeled in the present invention without any surrounding matter. Proteins almost always exist in aqueous solution; indeed, protein crystals contain between 20% and 90% water and dissolved salts which fill the space between the protein molecules. Certain kinds of amino acids have sidechains which make favorable interactions with aqueous solutions (serine, threonine, arginine, lysine, histidine, aspartic acid, glutamic acid, proline, asparagine, and glutamine) and are termed hydrophilic. Other amino acids have side chains which are apolar and make unfavorable interactions with water (phenylalanine, tryptophan, leucine, isoleucine, valine, methionine, and tyrosine) and are termed hydrophobic. In natural proteins, hydrophilic amino acids are almost always found on the surface, in contact with solvent; hydrophobic amino acids are almost always inside the protein in contact with other hydrophobic amino acids.

The remaining amino acids (alanine, glycine, and cysteine) are found both inside proteins and on their surfaces. The designs of the present invention should resemble natural proteins as much as possible, so hydrophobic residues are placed inside and hydrophilic residues are placed outside as much as possible.

Programs could be utilized to calculate an energy for each hypothetical structure. In addition, programs could make local adjustments to the hypothetical molecules to minimize the energy. Finally, molecular dynamics could be used to identify particularly unstable parts of the hypothetical molecule. Although existing programs could calculate a nominal energy for each hypothetical structure, it has not yet been demonstrated that such calculations can differentiate between sequences which will fold and those that will not. Energy minimization could also be accomplished with extant programs, but energy minimization also can not differentiate between sequences which will fold and those that will not. Molecular dynamics simulations currently cannot be continued long enough to simulate the actual folding or unfolding of a protein and so cannot distinguish between stable and unstable molecules.

Two megabytes of storage 128' in the computer generated display system 116 is added so that several different molecules can be stored at the display level. These molecules then can be switched back and forth on the color display 120 so that the expert operator can sequentially view them while making expert decisions. The parallel interface that is shown in Figure 2 would allow the coordinates to be trans-

ferred faster from the CPU 102' to the electronics stage 118' of the computer generated display system 116.

The parallel processing architecture embodiment of the present invention is described below in Section V. This parallel architecture embodiment provides even faster analysis and display.

III. Single Linker Embodiment

This first embodiment of the present invention determines and displays possible chemical structures for using a single linker to convert the naturally aggregated but chemically separate heavy and light polypeptide chains into a single polypeptide chain which will fold into a three dimensional structure very similar to the original structure made of two polypeptide chains.

A. Plausible Site Selection

There are two main goals of the plausible site selection step 302 of the present invention shown in very generalized block diagram form in Figure 3. The first goal is to select a first plausible site on the first chain that is the minimum distance from the second plausible site on the second chain. The first point on the first chain and the second point on the second chain comprise the plausible site.

The second goal of the site selection is to select plausible sites that will result in the least loss of native protein. Native protein is the original protein composed of the two aggregated polypeptide chains of the variable region. It is not chemically possible to convert two chains to one without altering some of

the amino acids. Even if only one amino acid was added between the carboxy terminal of the first domain and the amino terminal of the second domain, the charges normally present at these termini would be lost. In the variable regions of antibodies, the terminii of the H and L chains are not very close together. Hypothetical linkers which join the carboxy terminus of one chain to the amino terminus of the other do not resemble the natural variable region structures. Although such structures are not impossible, it is more reasonable to cut away small parts of the native protein so that compact linkers which resemble the native protein will span the gap. Many natural proteins are known to retain their structure when one or more residues are removed from either end.

In the present embodiment, only a single linker (amino acid sequence or bridge for bridging or linking the two plausible sites to form a single polypeptide chain) is used. Figure 4 shows in block diagram form the steps used to select plausible sites in the single linker. The steps of Figure 4 are a preferred embodiment of step 302 of Figure 3.

A domain 1 is picked in a step 402 (see Figure 4). A schematic diagram of two naturally aggregated but chemically separate polypeptide chains is shown in Figure 5A. For purposes of illustration, assume that L is the light chain of the antibody variable region (the first polypeptide chain) and is domain 1. As shown in Figure 5A, light chain L is on the left side, and heavy chain H is on the right side.

The next step 404 is to pick the domain 2, which, as indicated, is the heavy chain H of the antibody variable region on the right side of Figure 5A.

The linker that will be selected will go from domain 1 (the light chain L) towards domain 2 (heavy chain, H). As the linker will become part of the single polypeptide chain, it must have the same directionality as the polypeptides it is linking; i.e. the amino end of the linker must join the carboxy terminal of some amino acid in domain 1, and the carboxy terminal of the linker must join the amino terminal of some residue in domain 2. A starting point (first site) on domain 1 is selected, as represented by step in 406 in Figure 4. The starting point is chosen to be close to the C (C for carboxy) terminal of domain 1, call this amino acid tau 1. It is important to pick tau 1 close to the C terminal to minimize loss of native protein structure. Residue tau 1 is shown schematically in two dimensions in figure 6A; it is also shown in figure 6B where it is presented in a two-dimensional representation of the naturally aggregated but chemically separate H and L polypeptide chains.

Next, the final point (second site) close the N (N for amino) terminal of domain 2 is selected, as indicated by step 408 of Figure 4. The final site is an amino acid of domain 2 which will be called sigma 1. It is important that amino acid sigma 1 be close to the N terminal of domain 2 to minimize loss of native protein structure. Amino acid sigma 1 is shown schematically in figure 6A and in the more realistic representation of figure 6B.

Figure 7 shows in simplified form the concept that the linker goes from a first site at amino acid tau 1 in domain 1 to a second site at amino acid sigma 1 in

domain 2. There are a plurality of possible first sites and a plurality of second sites, as is shown in figure 7. A computer program prepares a table which contains for each amino acid in domain 1 the identity of the closest amino acid in domain 2 and the distance. This program uses the position of the alpha carbon as the position of the entire amino acid. The expert operator prepares a list of plausible amino acids in domain 1 to be the first site, tau 1, and a list of plausible amino acids in domain 2 to be the second site, sigma 1. Linkers are sought from all plausible sites tau 1 to all plausible sites sigma 1. The expert operator must exercise reasonable judgement in selecting the sites tau 1 and sigma 1 in deciding that certain amino acids are more important to the stability of the native protein than are other amino acids. Thus the operator may select sites which are not actually the closest.

The complete designed protein molecule in accordance with the present invention consists of the domain 1 (of the light chain L) up to the amino acid tau 1, the linker, as shown by the directional-line in Figure 8A and in Figure 8B, and the domain 2 from amino acid sigma 1 to the C terminus of the heavy chain, H. As shown in Figures 8A and 8B, in the representative example, this results in the following loss of native protein.

The first loss in native protein is from the residue after residue tau 1 to the C terminus of domain 1 (light chain L). The second loss of native protein is from the N terminus of domain 2 (heavy chain, H) to the amino acid before sigma 1.

As is best understood from Figure 8A, the introduction of linker 1 produces a single polypeptide chain from the two naturally aggregated chains. The polypeptide chain begins with the N terminal of domain 1. Referring now to Figure 8B, the chain proceeds through almost the entire course of the native light chain, L, until it reaches amino acid tau 1. The linker then connects the carboxy terminal of a very slightly truncated domain 1 to residue sigma 1 in the very slightly truncated domain 2. Since a minimum amount of native protein is eliminated, and the linker is selected to fit structurally as well as possible (as described below in connection with general steps 2 and 3 of the present invention), the resulting single polypeptide chain has a very high probability (several orders of magnitude greater than if the linker was selected randomly) to fold into a three-dimensional structure very similar to the original structure made of two polypeptide chains.

The single polypeptide chain results in a much more stable protein which contains a binding site very similar to the binding site of the original antibody. In this way a single polypeptide chain can be engineered from the naturally occuring two-polypeptide chain variable region, so as to create a polypeptide of only one chain, but maintaining the binding site of the antibody.

In the current mode of the present invention, the expert operator selects the sites with minimal help from the computer. The computer prepares the table of closest-residue-in-other-domain. The computer can provide more help in the following ways.

(1) Prepare a list of conserved and variable residues for variable regions of antibodies (Fv region). Residues which vary from Fv to Fv would be much better starting or ending sites for linkage than are residues which are conserved over many different Fv sequences.

(2) Prepare a list of solvent accessibilities. Amino acids exposed to solvent can be substituted with less likelihood of destabilizing the native structure than amino acids buried within the native structure. Exposed amino acids are better choices to start or end linkage.

With respect to each of the plurality of possible first sites (on domain 1 or light chain L) there are available a plurality of second sites (on domain 2 or heavy chain H) (See Figures 7 and 8A). As the second site is selected closer to the N terminus of domain 2, the distance to any of the plausible first sites increases. Also, as the first site is selected closer to the C terminus of domain 1 the distance to any of the plausible second sites increases. It is this tension between shortness of linker and retention of native protein which the expert operator resolves in choosing gaps to be linked. The penalty for including extra sites in the list of gaps are:

(1) searching in general step 2 will be slower; and

(2) more candidates will pass from step 2 many of which must be rejected in step 3. As step 3 is currently a manual step, this is the more serious penalty.

Figure 8B shows diagramatically by a directional arrow the possible links that can occur between the various

sites near the C terminal of domain 1 and the various sites near the N terminal of domain 2.

B. Selection of Candidates

In the second of the three general steps of the present invention as used in the single linker embodiment, plausible candidates for linking the site 1 on domain 1 with site 2 on domain 2 are selected from a much larger group of candidates. This process of winnowing out candidates results in the expert operator and/or expert system having a relatively small group of candidates to rank from most plausible to least plausible in the third general step of the present invention, as described in subsection C below.

Currently, there are approximately 250 protein structures, determined at 2.0 A or higher resolution, in the public domain. The structures of these very complicated molecules are determined using sophisticated scientific techniques such as X-ray crystallography, neutron diffraction, and nuclear magnetic resonance. Structure determination produces a file of data for each protein. The Brookhaven Protein Data Bank (BPDB) exemplifies a repository of protein structural information. Each file in BPDB contains many records of different types. These records carry the following information:

(1) Name of the protein and standard classification number,

(2) Organism from which protein was obtained,

(3) Name and address of contributor,

(4) Amino-acid sequence of each polypeptide chain, if known,

(5) Connectivity of disulfides, if any,

(6) Names and connectivities of any prosthetic groups, if any,

(7) References to literature,

(8) Transformation from reported coordinates to crystallographic coordinates,

(9) Coordinates of each atom determined.

There is at least one record for each atom for which a coordinate was determined. Some parts of some proteins are disordered and do not diffract X-rays, so no sensible coordinates can be given. Thus there may be amino acids in the sequence for which only some or none of the atoms have coordinates. Coordinates are given in Angstrom units (100,000,000 A = 1 cm) on a rectangular Cartesian grid. As some parts of a protein may adopt more than one spatial configuration, there may be two or more coordinates for some atoms. In such cases, fractional occupancies are given for each alternative position. Atoms move about, some more freely than others. X-ray data can give an estimate of atomic motion which is reported as a temperature (a.k.a. Debye-Waller) factor.

Any other data base which included, implicitly or explicitly, the following data would be equally useful:

(1) Amino acid sequence of each polypeptide chain.

(2) Connectivity of disulfides, if any,

(3) Names and connectivities of any prosthetic groups, if any,

(4) Coordinates (x, y, z) of each atom in each observed configuration.

(5) Fractional occupancy of each atom,

(6) Temperature factor of each atom.

Proteins usually exist in aqueous solution. Although protein coordinates are almost always determined for proteins in crystals, direct contacts between proteins are quite rare. protein crystals contain from 20% to 90% water by volume. Thus one usually assumes that the structure of the protein in solution will be the same as that in the crystal. It is now generally accepted that the solution structure of a protein will differ from the crystal structure only in minor details. Thus, given the coordinates of the atoms, one can calculate quite easily the solvent accessibility of each atom.

In addition, the coordinates implicitly give the charge distribution throughout the protein. This is of use in estimating whether a hypothetical molecule (made of native protein and one or more linkers) will fold as designed. The typical protein whose structure is known comprises a chain of amino acids (there are 21 types of amino acids) in the range of 100 to 300 amino acids.

Each of these amino acids alone or in combination with the other amino acids as found in the known protein molecule can be used as a fragment to bridge the two sites. The reason that known protein molecules are used is to be able to use known protein fragments for the linker or bridge.

Even with only 250 proteins of known structure, the number of possible known fragments is very large. A linker can be from one to twenty or thirty amino acids long. Let "Lmax" be the maximum number of amino acids allowed in a linker, for example, Lmax might be

25. Consider a protein of "Naa" amino acids. Proteins have Naa in the range 100 to 800, 250 is typical. From this protein one can select Naa-1 distinct two-amino-acid linkers, Naa-2 distinct three-amino-acid linkers,...and (Naa+1-Lmax) distinct linkers containing exactly Lmax amino acids. The total number of linkers containing Lmax or fewer linkers is "Nlink,"

$$\mathrm{Nlink} = \sum_{j=1,\mathrm{Lmax}} (\mathrm{Naa}+1-j)$$

$$= \mathrm{Naa} \times (\mathrm{Lmax}) - (\mathrm{Lmax} \times \mathrm{Lmax})/2 + \mathrm{Lmax}/2$$

If Naa is 250 and Lmax is 25, Nlink will be 5975. If the number of known proteins is "Nprot," then the total number of linkers, "Nlink_total" will be

$$\mathrm{Nlink\_total} = \sum_{k=1,\mathrm{Nprot}} \quad \sum_{j=1,\,\mathrm{Lmax}} (\mathrm{Naa}(k)+1-j)$$

$$= \sum_{k=1,\,\mathrm{Nprot}} [\mathrm{Naa}(k)\mathrm{x}(\mathrm{Lmax}) - (\mathrm{LmaxxLmax})/2+\mathrm{Lmax}/2]$$

$$= \mathrm{Nprotx}(\mathrm{Lmax}/2-\mathrm{Lmax} \times \mathrm{Lmax})/2 + \mathrm{Lmax} \times \sum_{K=1,\,\mathrm{Nprot}} \mathrm{Naa}(k)$$

Where Naa(k) is the number of amino acids in the kth protein. With 250 proteins, each containing 250 amino acids (on average), and Lmax set to 25, Nlink_total is 1,425,000.

This is the number of linkers of known structure. If one considers the number of possible amino acid sequences up to length Lmax (call it "Nlink_possible"), it is much larger.

$$\text{Nlink\_possible} = \sum_{J = 1, \text{Lmax}} 20^J$$

For Lmax = 25

Nlink_possible = 353, 204, 547, 368, 421, 052, 631, 578, 947, 368, 420

$$= 3.53 * 10^{32}$$

Using <u>known</u> peptide fragments thus reduces the possibilities by twenty-six orders of magnitude. Appropriate searching through the known peptide fragments reduces the possibilities a further five orders of magnitude.

Essentially, the present invention utilizes a selection strategy for reducing a list of possible candidates. This is done as explained below in a preferred form in a three step process. This three step process, as is illustrated in the explanation of the each of the three steps of the process, significantly reduces the computer time required to extract the most promising candidates from the data base of possible candidates. This should be contrasted with a serial search throughout the entire data base of candidates,

which would require all candidates to be examined in total. The present invention examines certain specific parameters of each candidate, and uses these parameters to produce subgroups of candidates that are then examined by using other parameters. In this way, the computer processing speed is significantly increased.

The best mode of the present invention uses a protein data base created and supplemented by the Brookhaven National Laboratory in Upton, Long Island, New York. This data base is called the Brookhaven Protein Data Base (BPDB). It provides the needed physical and chemical parameters that are needed by the present invention. It should be understood, that the candidate linkers can be taken from the Brookhaven Protein Data Base or any other source of three-dimensional protein structures. These sources must accurately represent the proteins. In the current embodiment, X-ray structures determined at resolution of 2.5A or higher and appropriately refined were used. Each peptide is replaced (by least-squares fit) by a standard planar peptide with standard bond lengths and angles. Peptides which do not accurately match a standard peptide (e.g. cis peptides) are not used to begin or end linkers, but may appear in the middle.

Each sequence up to some maximum number of amino acids (Lmax) is taken as a candidate. In the preferred embodiment, the maximum number of amino acids (Lmax) is set to 30. However, the present invention is not limited to this number, but can use any maximum number that is desired under the protein engineering circumstances involved.

1. <u>Selecting Candidates with Proper Distance Between the N Terminal and the C Terminal</u>.

The first step in the selection of candidates step is to select the candidate linkers with a proper distance between the N terminal and the C terminal from all of the candidate linkers that exist in the protein data base that is being used. Figure 9 shows in block diagram form the steps that make up this candidate selection process utilizing distance as the selection parameter.

Referring to Figure 9, a standard point relative to the peptide unit at the first site is selected, as shown by block 902.

A standard point relative to the peptide unit in the second site is also picked, as indicated by a block 904. Note that in the best mode the geometric centers of the peptide units of the first and second sites are used, but any other standard point can be utilized, if desired.

The distance between the standard points of the two peptides at the first and second sites defining the gap to be bridged by the linker is then calculated, as indicated by block 906. This scalar distance value is called the Span of the gap. Note that this scalar value does not include any directional information.

Next, as indicated by a step 908, the distance between the ends of the possible linker candidates are calculated. The distance between the ends of a particular candidate is called the span of the candidate. Note that each possible linker candidate has a span of the candidate scalar value.

The final step in the distance selection candidate selection process is that of a step 910. In step 910, candidates are discarded whose span of the candidate values differ from the span of the gap value by more than a preselected amount (this preselected amount is Max LSQFIT error). In the best mode of the present invention, the preselected amount for Max LSQFIT error is 0.50 Angstroms. However, any other suitable value can be used.

The preceding discussion has been for a single gap. In fact, the expert user often selects several gaps and the search uses all of them. The span of each candidate is compared to the span of each gap until it matches one, within the preset tolerance, or the list of gaps is exhausted. If the candidate matches none of the gaps, it is discarded. If it matches any gap it is carried to the next stage.

The inventors have determined that the use of the distance as the first parameter for discarding possible linker candidates results in a significant reduction in the number of possible candidates with a minimum amount of computer time that is needed. In terms of the amount of reduction, a representative example (using linkers up to 20 amino acids) starts out with 761,905 possible candidates that are in the protein data base. This selection of candidates using the proper distance parameter winnows this number down to approximately 63,727 possible candidates. As is discussed below, the distance selection operation requires much less computer time than is required by the other two steps which make up this selection step 304.

The result of this selection of candidates according to proper distance is a group (called a first group of candidates) which exhibit a proper length as compared to the gap that is to be bridged or linked. This first group of candidates is derived from the protein data base using the distance criteria only.

2. Selecting Candidates with Proper Direction from N Terminal to C Terminal

This substep essentially creates a second group of possible candidates from the first group of possible candidates which was produced by the distance selection substep discussed in connection with Figure 9. The second group of candidates is selected in accordance with the orientation of the C terminal residue (i.e. the final residue) of the linker with respect to the N terminal residue (i.e. the initial residue) which is compared to the orientation of the C terminal residue (i.e. the second site) of the gap with respect to the N terminal residue (i.e. the first site). See Figure 20B. In this way, this direction evaluation determines if the chain of the linker ends near the second site of the gap, when the amino terminal amino acid of the linker is superimposed on the first site of the gap so as to produce the minimum amount of unwanted molecular distortion.

Referring now to Figure 10, the first step used in producing the second group of possible candidates is a step 1002. In step 1002 a local coordinate system is established on the N terminal residue of one of the selected gaps. For example, one might take the local X-axis as running from the first alpha carbon of the N

terminal residue to the second alpha carbon of the N terminal residue, with the first alpha carbon at the origin · the second alpha carbon on the plus X-axis. The local Y-axis is selected so that the carbonyl oxygen lies in the xy plane with a positive y coordinate. The local Z-axis is generated by crossing X into Y. Next, as indicated by step 1004, a standard reference point in the C terminal residue of the gap is located and its spherical polar coordinates are calculated in the <u>local</u> system. The standard reference point could be any of the atoms in the C terminal peptide (throughout this application, peptide, residue, and amino acid are used interchangeably) or an average of their positions. Steps 1002 and 1004 are repeated for all gaps in the list of gaps. As indicated by step 1006, a local coordinate system is established on the N terminal residue of one of the candidates. This local coordinate system must be established in the same manner used for the local coordinate systems established on each of the gaps. Various local systems could be used, but one must use the same definition throughout. In step 1008, the standard reference point is found in the C terminal residue of the current candidate. This standard point must be chosen in the same manner used for the gaps. The spherical polar coordinates of the standard point are calculated in the local system of the candidate. (This use of local coordinate system is completely equivalent to rotating and translating all gaps and all candidates so that their initial peptide lies in a standard position at the origin.) In step 1010, the spherical polar coordinates of the gap vector (r, theta, phi) are compared

to the spherical polar coordinates of the candidate vector (r, theta, phi). In step 1012 a preset threshhold is applied, if the two vectors agree closely enough, then one proceeds to step 1014 and enrolls the candidate in the second group of candidates. Currently, this preset threshhold is set to 0.5 A, but other values could be used. From step 1014, one skips forward to step 1022, vide infra. On the other hand, if the vectors compared in step 1012 are not close enough, one moves to the next gap vector in the list, in step 1016. If there are no more gaps, one goes to step 1018 where the candidate is rejected. If there are more gaps, step 1020 increments the gap counter and one returns to step 1010. From steps 1014 or 1018 one comes to step 1022 where one tests to see if all candidates have been examined. If not, step 1024 increments the candidate counter and one returns to step 1006. If all candidates have been examined, one has finished, step 1026.

Figure 11 shows the concept of comparing the direction of the gap to the direction of the candidate.

The inventors have determined that in the example discussed above where 761,905 possible candidates are in the protein data base, the winnowing process in this step reduces the approximate 63,727 candidates in the first group to approximately 50 candidates in the second group. The inventors have also determined that as referenced to the units of computer time referred to above in connection with the scalar distance parameter, it takes approximately 4 to 5 computer units of time to perform the selection of this step. Thus, it can be appreciated that it preserves computer time to

perf rm the distance selection first, and the direction selection second since the direction selection process takes more time than the distance selection process.

3. <u>Selecting Candidates with Proper Orientation at Both Termini</u>

In this step, the candidates in the second group of step 1016 of Figure 10 are winnowed down to produce a third group of plausible candidates using an evaluation of the relative orientation between the peptide groups at either end of the candidate, compared to the relative orientation between the peptide groups at either end of the gap. In a step 1201, (Figure 12) decide that a peptide will be represented by 3, 4, or 5 atoms (<u>vide infra</u>). Specifically, in a step 1202, one of the candidates in the second group (step 1014) is selected for testing. In a step 1204, three to five atoms in the first peptide are selected to define the orientation of the first peptide. So long as the atoms are not collinear, three atoms is enough, but using four or five atoms makes the least-squares procedure which follows over-determined and therefore compensates for errors in the coordinates. For example, assume selection of four atoms: C <u>alpha</u>, C, N, and C <u>beta</u>. Next, in a step 1206, one selects the corresponding 3,4, or 5 atoms from the final peptide of the selected candidate. These 6, 8, or 10 atoms define a three-dimensional object. In a step 1208, select one of the gaps. Select the corresponding 6, 8, or 10 atoms from the gap. In a step 1210, least-squares fit the atoms from the candidate to the atoms from the gap. This least-squares fit allows degrees

of freedom to superimpose the two three-dimensional objects. Assume that one object is fixed and the other is free to move. Three degrees of freedom control the movement of the center of the free object. Three other degrees of freedom control the orientation of the free object. In a step 1212, the result of the least-square fit is examined. If the Root-Mean-Square (RMS) error is less than some preset threshhold, the the candidate is a good fit for the gap being considered and is enrolled in the third group in a step 1214. If, on the other hand, the RMS error is greater than the preset threshhold, one checks to see if there is another gap in the list in a step 1216. If there is, one selects the next gap and returns to step 1208. If there are no more gaps in the list, then the current candidate from the second group is rejected in step 1218. In step 1220, one checks to see if there are more candidates in the second group; if so, a new candidate is selected and one returns to step 1201. If there are no more candidates, one is finished (step 1222). Again referring to a representative case, where linkers of length up to twenty amino acids were sought for a single gap with separation 12.7 A, the protein data bank contained 761,905 potential linkers. Of these, 63,727 passed the distance test. The direction test removed all but 50 candidates. The orientation test passed only 1 candidate with RMS error less than or equal to 0.5 A. There were two additional candidates with RMS error between 0.5 A and 0.6 A. Moreover, the inventors have determined that it takes about 25 units of computer time to evaluate each candidate in group 2 to decide whether they should be

selected for group 3. It can be appreciated now that the order selected by the inventors for the three steps of winnowing the candidates has been selected so that the early steps take less time per candidate than the following steps. The order of the steps used to select the candidate can be changed, however, and still produce the desired winnowing process. Logically, one might even omit steps one and two and pass all candidates through the least-squares process depicted in Figure 12 and achieve the same list of candidates, but at greater cost in computing. This may be done in the case of parallel processing where computer time is plentiful, but memory is in short supply.

Another approach (not illustrated) for determining whether the proper orientation exists between the ends of the candidate, is to examine only the atoms at the C terminal of the candidate as compared to the atoms at the final peptide of the gap. In step 2, the inventors aligned the first peptide of the candidate with the first peptide in the gap. Having done this, one could merely compare the atoms at the C terminal of the candidate with the atoms of the second peptide of the gap. This approach is inferior to that discussed above because all the error appears at the C terminus, while the least-squares method discussed above distributes the errors evenly.

C. <u>Ranking and Eliminating Candidates</u>.

As shown in Figure 3, the third general step in the present invention is that of ranking the plausible candidates from most plausible to least plausible, and eliminating those candidates that do not appear to be plausible based on criteria utilized by an expert operator and/or expert system.

In the best mode, the candidates in the third group (step 1214) are provided to the expert operator, who can sequentially display them in three dimensions utilizing the computer-graphics display system 116. The expert operator then can make decisions about the candidates based on knowledge concerning protein chemistry and the physical relationship of the plausible candidate with respect to the gap being bridged. This analysis can be used to rank the plausible candidates in the third group from most plausible to least plausible. Based on these rankings, the most plausible candidates can be selected for genetic engineering.

As noted above in connection with the illustrative example, there are typically few (under 100) candidates which make it to the third group of step 1214. Consequently, a moderately expert operator (one having a Bachelor of Science degree in chemistry, for example), can typically winnow down this number of plausible candidates to a group of 10 to 15. Thereafter, a more expert operator and/or expert system can further winnow down the number. In this way, only a very few of the plausible candidates needs to be tested in practice as compared to the hundreds, thousands or more of candidates that would have to be tested if no selection process like that of the present invention was used. This speeds up the process of engineering the single chain molecules by orders of magnitude, while reducing costs and other detriments by orders of magnitude as well.

In certain situations, however, automatic ranking in this third general step may be warranted. This could occur, for example, where the expert operator

was presented with quite a few candidates in the third group, or where it is desired to assist the expert operator in making the ranking selections and eliminating candidates based on prior experience that has been derived from previous engineering activities and/or actual genetic engineering experiments.

Referring now to Figure 13, a coordinate listing of the hypothetical molecule (candidate) is automatically constructed, as is indicated by a block 1302. The expert operator can then display using a first color the residues from domain 1 of the native protein. Color display 120 can provide a visual indication to the expert operator of where the residues lie in domain 1. This is indicated by a block 1304.

The expert operator then can display on color display 120 the residues from domain 2 of the native protein using a second color, as is indicated by a block 1306. The use of a second color provides a visual indication to the user which assists in distinguishing the residues from domain 1 from the residues from domain 2.

The linker (candidate) being ranked can be displayed in a selected color, which color can be different from the first color of step 1304 and/or the second color from step 1306. Again, by using this visual color indication, the expert operator can distinguish the residues of domain 1 and 2 of the native protein. This display of the linker candidate is indicated by a block 1308.

The initial picture on the color display 120 provided to the expert operator typically shows the alpha carbons for all of the residues. This is indicated by

a block 1310. In addition, the initial picture shows the main-chain and side-chains for residues and linkers and one residue before the linker and one residue after the linker. This is indicated by a block 1312.

The expert operator can also cause any of the other atoms in the native protein or linker candidate to be drawn at will. The molecule can be rotated, translated, and enlarged or reduced, by operator command, as was discussed generally in connection with the computer-graphics display system 116 above. The block diagram of Figure 13 indicates that each of the steps just discussed are accomplished in serial fashion. However, this is only for purposes of illustration. It should be understood that the operator can accomplish any one or more of these steps as well as other steps at will and in any sequence that is desired in connection with the ranking of the plausible candidates in group 3.

The expert operator and/or expert system utilized in this third general step in ranking the candidates from most plausible to least plausible and in eliminating the remaining candidates from group 3, can use a number of different rules or guidelines in this selection process. Representive of these rules and guidelines are the following which are discussed in connection with Figure 14. Note that the blocks in Figure 14 show the various rules and/or criteria, which are not necessarily utilized in the order in which the boxes appear. The order shown is only for purposes of illustration. Other rules and/or criteria can be utilized in the ranking process, as well.

As shown in step 1402, a candidate can be rejected if any atom of the linker comes closer than a minimum allowed separation to any retained atom of the native protein structure. In the best mode, the minimum allowed separation is set at 2.0 Angstroms. Note that any other value can be selected. This step can be automated, if desired, so that the expert operator does not have to manually perform this elimination process.

A candidate can be penalized if the hydrophobic residues have high exposure to solvent, as is indicated by a block 1404. The side chains of phenylananine, tryptophan, tyrosine, leucine, isoleucine, methionine, and valine do not interact favorably with water and are called hydrophobic. Proteins normally exist in saline aqueous solution; the solvent consists of polar molecules (H_2O) and ions.

A candidate can be penalized when the hydrophilic residues have low exposure to solvent. The side chains of serine, threonine, aspartic acid, glutamic acid, asparagine, glutamine, lysine, arginine, and proline do interact favorably with water and are called hydrophilic. This penalization step for hydrophilic residues is indicated by a block 1406.

A candidate can be promoted when hydrophobic residues have low exposure to solvent, as is indicated by a block 1408.

A candidate can be promoted when hydrophilic residues have high exposure to solvent, as indicated by a block 1410.

A candidate can be penalized when the main chain fails to form hydrogen bonds, as is indicated by a block 1412.

A candidate can be penalized when the main chain makes useless excursions into the solvent region. Useless excursions are those which do not make any evident interaction with the retained native protein. This is indicated by a block 1414.

A candidate can be promoted when the main chain forms a helix, as is indicated by a block 1416. Hilices are self-stabilizing. Thus a linker which is helical will be more stable because its main-chain polar atoms (O and N) will form hydrogen bonds within the linker.

As is indicated by a block 1418, a candidate can be promoted when the main chain forms a beta sheet which fits against existing beta sheets. The strands of beta sheets stabilize each other. If a linker were found which was in a beta-sheet conformation such that it would extend an existing beta sheet, this interaction would stabilize both the linker and the native protein.

Another expert design rule penalizes candidates which have sterically bulky side chains at undesirable positions along the main chain. Furthermore, it is possible to "save" a candidate with a bulky side chain by replacing the bulky side chain by a less bulky one. For example if a side chain carries a bulky substituent such as leucine or isoleucine, a possible design step replaces this amino acid by a glycine, which is the least bulky side chain.

Other rules and/or criteria can be utilized in the selection process of the third general step 306, and the present invention is not limited to the rules and/or criteria discussed. For example, once the

linker has been selected it is also possible to add, delete, or as stated, modify one or more amino acids therein, in order to accomplish an even better 3-D fit.

IV. Double and Multiple Linker Embodiments

Section III above described the single linker embodiment in accordance with the present invention. This section describes double linker and multiple linker embodiments in accordance with the present invention. For brevity purposes, only the significant differences between this embodiment and the single linker embodiment will be described here and/or illustrated in separate figures. Reference should therefore be made to the text and figures that are associated with the single linker embodiment

A. Plausible Site Selection.

The two main goals of minimizing distance between the sites to be linked and the least loss of native protein apply in the site selection in the double and multiple linker embodiments as they did apply in the single linker embodiment discussed above.

Figure 15A shows a simplified two dimensional representation of the use of two linkers to create the single polypeptide chain from the two naturally aggregated but chemically separate polypeptide chains. Figure 15B shows in two dimensions a three dimensional representation of the two chains of Figure 15A. Referring now to Figures 15A and B, the first step in de-

termining suitable sites is to find a site in domain 1 which is close to either the C or N terminus of domain 2. For purposes of illustration, and as is shown in Figures 15A and 15B, it is assumed that the most promising location is the C terminus of domain 2. The residue in domain 1 is called <u>Tau</u> 1, while the residue in domain 2 is called <u>Sigma</u> 1.

Figures 16A and 16B are respectively two dimensional simplified plots of the two chains, and two dimensional plots of the three dimensional representation of the two chains. They are used in connection with the explanation of how plausible sites are selected for the second linker in the example situation.

The first step in connection with finding plausible sites for the second linker is to find a residue in domain 1 that is before <u>Tau</u> 1 in the light chain. This residue is called residue <u>Tau</u> 2. It is shown in the top portion in Figure 16A, and in the right middle portion in Figure 16B.

The next step in the site selection process for the second linker is to find a residue in domain 2 near the N terminus of domain 2. This residue is called residue <u>Sigma</u> 2. Reference again is made to Figures 16A and B to show the location of <u>Sigma</u> 2.

The second linker (linker 2) thus runs from <u>Tau</u> 2 to <u>Sigma</u> 2. This is shown in Figures 17A and 17B. Note that the chain that is formed by these two linkers has the proper direction throughout.

Figure 18 shows in two dimensional simplified form the single polypeptide chain that has been formed by the linking of the two independent chains using the two linkers. Note that the approach outlined above

resulted in the minimal loss of native protein. The completely designed protein is shown in Figure 17 and consists of domain 1 from the N terminal to Tau 2, linker 2, domain 2 from Sigma 2 to Sigma 1, linker 1, and domain 1 from Taul to the C terminus. The arrows that are shown in Figure 17 indicate the direction of the chain.

Figure 17 shows that the residues lost by the utilization of the two linkers are: (a) from the N terminus of domain 2 up to the residue before Sigma 2; and (b) from the residue after Sigma 1 to the C terminus of domain 2; and (c) from the residue after Tau 2 to the residue before Tau 1 of domain 1.

If one of the linkers in the two linker case is very long, one could link from Tau 2 to a residue in domain 2 after Sigma 1. A third linker (not shown) would then be sought from a residue near the C terminal of domain 2 to a residue near the N terminal of domain 2.

Additionally, one could use two linkers to reconnect one of the domains in such a way that a single linker or a pair of linkers would weld the two domains into one chain.

B. Candidate Selection and Candidate Rejection Steps

Ranking of linkers in the multilinker cases follows the same steps as in the single linker case except there are some additional considerations.

(1) There may be a plurality of linkers for each of the two (or more) gaps to be closed. One must consider all combinations of each of the linkers for gap A with each of the linkers for gap B.

(2) One must consider the interactions between linkers.

As one must consider combinations of linkers, the ranking of individual linkers is used to cut down to a small number of very promising linkers for each gap. If one has only three candidates for each gap, there are nine possible constructs.

The process of examining interactions between linkers and discarding poor candidates can be automated by applying the rules discussed above.

V. Parallel Processing Embodiment

Figure 19 shows in block diagram form the parallel processing approach that can be utilized in the present invention.

As shown in Figure 19, a friendly serial processor 1902 is connected by a first bus 1904 to a plurality of data storage devices and input devices. Specifically, and only for purposes of illustration, a tape input stage 1906 is connected to bus 1904 so as to read into the system the parameters of the protein data base that is used. A high storage disk drive system 1908 (having, for example, 5 gigabits of storage) is also connected to bus 1904. Operationally, for even larger storage capabilities, an optical disk storage stage 1910 of conventional design can be connected to bus 1904.

The goal of the hypercube 1912 that is connected to the friendly serial processor 1902 via a bi-directional bus 1914 is twofold: to perform searching faster, and to throw out candidates more automatically.

The hypercube 1912, having for example, 2^{10} to 2^{16} nodes provides for parallel processing. There are computers currently available which have up to 1,024 computing nodes. Thus each node would need to hold only about 1400 candidate linkers and local memory of available machines would be sufficient. This is the concept of the hypercube 1912. Using the hypercube parallel processing approach, the protein data base can be divided into as many parts as there are computing nodes. Each node is assigned to a particular known protein structure.

The geometry of the gap that has to be bridged by a linker is sent by the friendly serial processor 1902 via bus 1914 to the hypercube stage 1912. Each of the nodes in the hypercube 1912 then processes the geometrical parameters with respect to the particular candidate linker to which it is assigned. Thus, all of the candidates can be examined in a parallel fashion, as opposed to the serial fashion that is done in the present mode of the present invention. This results in much faster location (the inventors believe that the processing speed can be brought down from 6 hours to 3 minutes using conventional technology) in locating the candidates that can be evaluated by the second step 304 of the present invention.

Another advantage for the parallel processing embodiment is that it will provide sufficient speed to allow candidates to be thrown out more automatically. This would be achieved using molecular dynamics and energy minimization. While this could be done currently on serial processing computers (of the super computer variety such as those manufactured by Cray

and Cyber) the parallel processing approach will perform the molecular dynamics and energy minimization much faster and cheaper than using the super computing approach.

In particular, hypercube computers exist which have inexpensive computing nodes which compare very favorably to supercomputers for scalar arithmetic. Molecular dynamics and energy minimization are only partly vectorizable because the potential functions used have numerous data-dependent branches.

VI. Preparation and Expression of Genetic Sequences, and Uses.

The polypeptide sequences generated by the methods described herein, give rise by application of the genetic code, to genetic sequences coding therefor. Given the degeneracy of the code, however, there are in many instances multiple possible codons for any one amino acid. Therefore, codon usage rules, which are also well understood by those of skill in the art, can be utilized for the preparation of optimized genetic sequences for coding in any desired organism. (See, for example, Ikemura, J. Mol. Biol. 151:389-409 (1981)).

Generally, it is possible to utilize the cDNA sequences obtained from the light and heavy chains of the variable region of the original antibody as a starting point. These sequences can then be joined by means of genetic linkers coding for the peptide linker candidates elucidated by the methods of the invention. The genetic sequence can be entirely synthesized *de novo* or fragments of cDNA can be linked together with the synthetic linkers, as described.

A large source of hybridomas and their corresponding monoclonal antibodies are available for the preparation of sequences coding for the H and L chains of the variable region. As indicated previously, it is well known that most "variable" regions of antibodies of a given class are in fact quite constant in their three dimensional folding pattern, except for certain specific hypervariable loops. Thus, in order to choose and determine the specific binding specificity of the single chain binding protein of the invention it becomes necessary only to define the protein sequence (and thus the underlying genetic sequence) of the hypervariable region. The hypervariable region will vary from binding molecule to molecule, but the remaining domains of the variable region will remain constant for a given class of antibody.

Source mRNA can be obtained from a wide range of hybridomas. See for example the catalogue <u>ATCC Cell Lines and Hybridomas</u>, December 1984, American Type Culture Collection, 20309 Parklawn Drive, Rockville, Maryland 20852, U.S.A., at pages 5-9. Hybridomas secreting monoclonal antibodies reactive with a wide variety of antigens are listed therein, are available from the collection, and usable in the invention. Of particular interest are hybridomas secreting antibodies which are reactive with viral antigens, tumor associated antigens, lymphocyte antigens, and the like. These cell lines and others of similar nature can be utilized to copy mRNA coding for the variable region or determine amino acid sequence from the monoclonal antibody itself. The specificity of the antibody to be engineered will be determined by the original se-

lection process. The class of antibody can be determined by criteria known to those skilled in the art. If the class is one for which there is a three-dimensional structure, one needs only to replace the sequences of the hyper-variable regions (or complementary determining regions). The replacement sequences will be derived from either the amino acid sequence or the nucleotide sequence of DNA copies of the mRNA.

It is to be specifically noted that it is not necessary to crystallize and determine the 3-D structure of each variable region prior to applying the method of the invention. As only the hypervariable loops change drastically from variable region to variable region (the remainder being constant in the 3-D structure of the variable region of antibodies of a given class), it is possible to generate many single chain 3-D structures from structures already known or to be determined for each class of antibody.

For example, linkers generated in the Examples in this application (e.g., TRY40, TRY61 or TRY59, see below) are for Fv regions of antibodies of the IgA class. They can be used universally for any antibody, having any desired specificity, especially if the antibody is of the IgA class.

Expression vehicles for production of the molecules of the invention include plasmids or other vectors. In general, such vectors containing replicon and control sequences which are derived from species compatible with a host cell are used in connection with the host. The vector ordinarily carries a replicon site, as well as specific genes which are capable of providing phenotypic selection in transformed cells. For example, E. coli is readily transformed

using pBR322, a plasmid derived from an E. coli species. pBR322 contains genes for ampicillin and tetracycline resistance, and thus provides easy means for identifying transformed cells. The pBR322 plasmid or other microbial plasmids must also contain, or be modified to contain, promoters which can be used by the microbial organism for expression of its own proteins. Those promoters most commonly used in recombinant DNA construction include the beta lactamase, lactose promoter systems, lambda phage promoters, and the tryptophan promoter systems. While these are the most commonly used, other microbial promoters have been discovered and can be utilized.

For example, a genetic construct for a single chain binding protein can be placed under the control of the leftward promoter of bacteriophage lambda. This promoter is one of the strongest known promoters which can be controlled. Control is exerted by the lambda repressor, and adjacent restriction sites are known.

The expression of the single chain antibody can also be placed under control of other regulatory sequences which may be homologous to the organism in its untransformed state. For example, lactose dependent E. coli chromosomal DNA comprises a lactose or lac operon which mediates lactose utilization by elaborating the enzyme beta-galactosidase. The lac control elements may be obtained from bacteriophage lambda plac5, which is infective for E. coli. The lac promoter-operator system can be induced by IPTG.

Other promoter/operator systems or portions thereof can be employed as well. For example, colicin El,

galactose, alkaline phosphatase, tryptophan, xylose, tac, and the like can be used.

Of particular interest is the use of the O_L/P_R hybrid lambda promoter (see for example U.S. patent application Serial Number 534,982 filed September 3, 1983, and herein incorporated by reference).

Other preferred hosts are mammalian cells, grown *in vitro* in tissue culture, or *in vivo* in animals. Mammalian cells provide post translational modifications to immunoglobulin protein molecules including correct folding or glycosylation at correct sites.

Mammalian cells which may be useful as hosts include cells of fibroblast origin such as VERO or CHO-K1, or cells of lymphoid origin, such as the hybridoma SP2/0-AG14 or the myeloma P3x63Sg8, and their derivatives.

Several possible vector systems are available for the expression of cloned single chain binding proteins in mammalian cells. One class of vectors utilizes DNA elements which provide autonomously replicating extrachromosomal plasmids, derived from animal viruses such as bovine papilloma virus, polyoma virus, or SV40 virus. A second class of vectors relies upon the integration of the desired gene sequences into the host cell chromosome. Cells which have stably integrated the introduced DNA into their chromosomes can be selected by also introducing drug resistance genes such as *E. coli* GPT or Tn5neo. The selectable marker gene can either be directly linked to the DNA gene sequences to be expressed, or introduced into the same cell by co-transfection. Additional elements may also be needed for optimal synthesis of single chain binding

protein mRNA. These elements may include splice signals, as well as transcription promoters, enhancers, and termination signals. cDNA expression vectors incorporating such elements include those described by Okayama, H., Mol. Cel. Biol., 3:280 (1983), and others.

Another preferred host is yeast. Yeast provides substantial advantages in that it can also carry out post translational peptide modifications including glycosylation. A number of recombinant DNA strategies exist which utilize strong promoter sequences and high copy number of plasmids which can be utilized for production of the desired proteins in yeast. Yeast recognizes leader sequences on cloned mammalian gene products, and secretes peptides bearing leader sequences (i.e., pre-peptides).

Any of a series of yeast gene expression systems incorporating promoter and termination elements from the actively expressed genes coding for glycolytic enzymes produced in large quantities when yeasts are grown in mediums rich in glucose can be utilized. Known glycolytic genes can also provide very efficient transcription control signals. For example, the promoter and terminator signals of the phosphoglycerate kinase gene can be utilized.

Once the strain carrying the single chain building molecule gene has been constructed, the same can also be subjected to mutagenesis techniques using, chemical agents or radiation, as is well known in the art. From the colonies thus obtained, it is possible to search for those producing binding molecules with increased binding affinity. In fact, if the first lin-

ker designed with the aid of the computer fails to produce an active molecule, the host strain containing the same can be mutagenized. Mutant molecules capable of binding antigen can then be screened by means of a routine assay.

The expressed and refolded single chain binding proteins of the invention can be labelled with detectable labels such as radioactive atoms, enzymes, biotin/avidin labels, chromophores, chemiluminescent labels, and the like for carrying out standard immunodiagnostic procedures. These procedures include competitive and immunometric (or sandwich) assays. These assays can be utilized for the detection of antigens in diagnostic samples. In competitive and/or sandwich assays, the binding proteins of the invention can also be immobilized on such insoluble solid phases as beads, test tubes, or other polymeric materials.

For imaging procedures, the binding molecules of the invention can be labelled with opacifying agents, such as NMR contrasting agents or X-ray contrasting agents. Methods of binding, labelling or imaging agents to proteins as well as binding the proteins to insoluble solid phases are well known in the art. The refolded protein can also be used for therapy when labelled or coupled to enzymes or toxins, and for purification of products, especially those produced by the biotechnology industry. The proteins can also be used in biosensors.

Having now generally described this invention the same will be better understood by reference to certain specific examples which are included for purposes of illustration and are not intended to be limiting unless otherwise specified.

EXAMPLES

In these experiments, the basic Fv 3-D structure used for the computer assisted design was that of the anti-phosphoryl choline myeloma antibody of the IgA class, MCPC-603. The X-ray structure of this antibody is publicly available from the Brookhaven data base.

The starting material for these examples was monoclonal antibody cell line 3C2 which produced a mouse anti-bovine growth hormone (BGH). This antibody is an IgG_1 with a gamma 1 heavy chain and kappa light chain. cDNA's for the heavy and light chain sequences were cloned and the DNA sequence determined. The nucleotide sequences and the translation of these sequences for the mature heavy and mature light chains are shown in Figures 21 and 22 respectively.

Plasmids which contain just the variable region of the heavy and light chain sequences were prepared. A ClaI site and an ATG initiation codon (ATCGATG) were introduced before the first codon of the mature sequences by site directed mutagenesis. A HindIII site and termination codon (TAAGCTT) were introduced after the codon 123 of the heavy chain and the codon 109 of the light chain. The plasmid containing the V_H sequences is pGX3772 and that containing the V_L is pGX3773 (Figure 23).

The examples below were constructed and produced by methods known to those skilled in the art.

EXAMPLE 1

A. Computer Design

A two-linker example (referred to as TRY 40) was designed by the following steps.

First, it was observed that light chains were much easier to make in E. coli than were heavy chains. It was thus decided to start with light chain. (In the future, one could certainly make examples which begin with heavy chain because there is a very similar contact between a turn in the heavy chain and the exit strand of the light chain.)

Refer to stereo Figure 30A, which shows the light and heavy domains of the Fv from MOPC-603 antibody; the constant domains are discarded. A line joining the alpha carbons of the light chain is above and dashed. The amino terminus of the light chain is to the back and at about 10 o'clock from the picture center and is labeled "N." At the right edge of the picture, at about 2 o'clock is an arrow showing the path toward the constant domain. Below the light chain is a line joining the alpha carbons of the heavy chain. The amino terminus of the heavy chain is toward the viewer at about 7 o'clock and is also labeled "N." At about 4:30, one sees an arrow showing the heavy chain path to its constant domain.

The antigen-binding site is to the left, about 9 o'clock and between the two loops which project to the right above (light chain) and below (heavy chain).

In addition to the alpha carbon traces, there are three segments in which all non-hydrogen atoms have been drawn. These strands are roughly parallel and from upper right to lower left. They are

(a) Proline 46 to Proline 50 of the light chain.

(b) Valine 111 to Glycine 113 of the heavy chain.

(c) Glutamic acid 1 to glycine 10 of the heavy chain.

The contact between tryptophan 112 of the heavy chain and proline 50 of the light chain seems very favorable. Thus it was decided that these two residues should be conserved. Several linkers were sought and found which would join a residue at or following Tryptophan 112 (heavy) to a residue at or following Proline 50 (light). Stereo figure 30B shows the region around TRP 112H in more detail. The letter "r" stands between the side-chain of TRP 112H and PRO 50L; it was wished to conserve this contact. The letter "q" labels the carboxy terminal strand which leads towards the constant domain. It is from this strand that a linker will be found which will connect to PRO 50L.

Once a linker is selected to connect 112H to 50L, one needs a linker to get from the first segment of the light chain into the beginning portion of the heavy chain. Note that PRO 46L turns the chain toward PRO 50L. This turning seemed very useful, so it was decided to keep PRO 46L. Thus the second linker had to begin after 46L and before 50L, in the stretch marked "s." A search for linkers was done beginning on any of the residues 46L, 47L, or 48L. Linkers beginning on residue 49L were not considered because the chain has already turned toward 50L and away from the amino terminal of the heavy chain. Linkers were sought which ended on any of the residues 1H to 10H.

Figure 30C shows the linked structure in detail. After TRP 112H and GLY 113H, was introduced the sequence PRO-GLY-SER, and then comes PRO 50L. A computer program was used to look for short contacts between atoms in the linker and atoms in the retained part of the Fv. There is one short contact between the beta carbon of the SER and PRO 50L, but small movements would relieve that. This first linker runs from the point labeled "x" to the point labeled "y." The second linker runs from "v" to "w." Note that most of the hydrophobic residues (ILE and VAL) are inside. There is a PHE on the outside. In addition, the two lysine residues and the asparagine residue are exposed to solvent as they ought to be. Figure 30D shows the overall molecule linked into a single chain.

B. Genetic Constructs

These constructs were prepared and the plasmids containing them using E. coli. hosts. Once constructed, the sequences can be inserted into whichever expression vehicle used in the organism of choice.

The first construction was TRY40 (the two-linker construction) which produces a protein with the following sequence:

Met-[L-chain 1-41]-Ile-Ala-Lys-Ala-Phe-Lys-Asn-[H-chain 8-105]-Pro-Gly-Ser-[L-chain 45-109]. The nucleotide sequence and its translation are seen in Figure 24. The hypervariable regions in TRY40 (as in TRY61 59 and 104B, see below) correspond, as indicated, to an IgG1 anti BGH antibody, even though the 3-D analysis was done on the Fv region of MCPC-603 antibody, having a different specificity, (anti phosphoryl choline) but having a similar framework in the variable region.

The antibody sequences in the plasmids pGX3772 and pGX3773 were joined to give the sequence of TRY40 in the following manner. The plasmids used contained an M13 bacteriophage origin of DNA replication. When hosts containing these plasmids are superinfected with bacteriophage M13 two types of progeny are produced, one containing the single-strand genome and the other containing a specific circular single-strand of the plasmid DNA. This DNA provided template for the oligonucleotide directed site specific mutagenesis experiments that follow. Template DNA was prepared from the two plasmids. An EcoRI site was introduced before codon 8 of the V_H sequence in pGX3772, by site directed mutagenesis, producing pGX3772'. Template from this construction was prepared and an XbaI site was introduced after codon 105 of the V_H sequence producing pGX3772''.

An EcoRI and an XbaI site were introduced into pGX3773 between codons 41 and 45 of the V_L sequence by site directed mutagenesis producing pGX3773'.

To begin the assembly of the linker sequences plasmid pGX3773' (V_L) DNA was cleaved with EcoRI and XbaI and treated with calf alkaline phosphatase. This DNA was ligated to the EcoRI to XbaI fragment purified from plasmid pGX3772''(V_H) which had been cleaved with the two restriction enzymes. The resulting plasmid pGX3774, contained the light and heavy chain sequences in the correct order linked by the EcoRI and XbaI restriction sites. To insert the correct linker sequences in frame, pGX3774 template DNA was prepared. The EcoRI junction was removed and the linker coding for the -Ile-Ala-Lys-Ala-Phe-Lys-Asn- inserted by site

directed mutagen sis, producing plasmid pGX3774'. Template DNA was prepared from this construction and the XbaI site corrected and the linker coding for -Pro-Gly-Ser- inserted by site directed mutagenesis producing plasmid pGX3775. The sequence was found to be correct as listed in Figure 24 by DNA sequencing.

In order to express the single-chain polypeptide, the sequence as a ClaI to HindIII fragment was inserted into a vector pGX3703. This placed the sequence under the control of the O_L/P_R hybrid lambda promoter (U.S. Patent Application 534,982, Sept. 23, 1983). The expression plasmid is pGX3776 (Figure 25). The plasmid pGX3776 was transformed into a host containing a heat sensitive lambda phage repressor; when grown at 30°C the synthesis of the TRY40 protein is repressed. Synthesis was induced by raising the temperature to 42°C, and incubating for 8-16 hours. The protein was produced at 7.2% of total cell protein, as estimated on polyacrylamide gel electropherograms stained with Coomassie blue.

EXAMPLE 2

A. Computer Design

A one-linker example (referred to as TRY 61) was designed by the following steps.

Refer to stereo Figure 31A which shows the light and heavy domains of the Fv; the constant domains are discarded. A line joining the alpha carbons of the light chain is dashed. The amino terminus of the light chain is to the back and at about the center of the picture and is labeled "N." At the right edge of the picture, at about 2 o'clock is an arrow showing

the path toward the constant domain of the light chain. Below the light chain is a line joining the alpha carbons of the heavy chain. The amino terminus of the heavy chain is toward the viewer at about 9 o'clock and is also labeled "N". At about 4:30, one sees an arrow showing the heavy chain path to its constant domain.

In addition to the alpha carbon traces, there are two segments in which all non-hydrogen atoms have been drawn. These segments are the last few residues in the light chain and the first ten in the heavy chain. Linkers were sought between all pairs of these residues, but only a few were found because these regions are widely separated.

Figure 31B shows the linker in place. Note that the molecule now proceeds from the amino terminal of the light chain to the carboxy terminal strand of the heavy chain. Note also that the antigen-binding region is to the left, on the other side of the molecule from the linker.

B. Genetic Constructs

The sequence of TRY61 (a single-linker embodiment) is Met-[L-chain 1-104]-Val-Arg-Gly-Ser-Pro-Ala-Ile-Asn-Val-Ala-Val-His-Val-Phe-[H-chain 7-123]. The nucleotide sequence and its translation are shown in Figure 26.

To construct TRY61, plasmid pGX3772' DNA was cleaved with *Cla*I and *Eco*RI and treated with calf alkaline phosphatase. This DNA was ligated with the *Cla*I to *Hind*III fragment from pGX3773 and two oligonucleotides which code for the linker sequence and have *Hind*III and *Eco*RI ends, so that the linker can

only be ligated in the correct orientation. Th resulting plasmid, pGX3777, was used to prepare template DNA. This DNA was used for site directed mutagenesis to remove the HindIII site inside the antibody sequences. The correct construction, pGX3777', was used to make template DNA for a site directed mutagenesis to remove the EcoRI site. The ClaI to HindIII fragment from the final construction, pGX3778, containing the TRY61 coding seouence was confirmed by DNA sequencing. The ClaI to HindIII was inserted into the pGX3703 expression vector. This plasmid is called pGX4904 (Figure 27). This plasmid was transformed into an E. coli host. The strain containing this plasmid has been induced, and the single chain protein produced as >2% of total cell protein.

EXAMPLE 3

A. Computer Design

A one-linker example (referred to as TRY 59) was designed by the following steps.

Refer to stereo Figure 32A which shows the light and heavy domains of the Fv; the constant domains are discarded. A line joining the alpha carbons of the light chain is above and dashed. The amino terminus of the light chain is to the back and at about 10 o'clock from the center of the picture and is labeled "N". At the right edge of the picture, at about 2 o'clock is an arrow showing the path toward the constant domain of the light chain. Below the light chain is a line joining the alpha carbons of the heavy chain. The amino terminus of the heavy chain is toward the viewer at about 8 o'clock and is also labeled

"N". At about 4:30, one sees an arrow showing the heavy chain path to its constant domain.

In addition to the alpha carbon traces, there are two segments in which all non-hydrogen atoms have been drawn. These segments are the last few residues in the light chain and the first ten in the heavy chain. Linkers we sought between all pairs of these residues, but only a few were found because these regions are widely separated.

Figure 32B shows the linker in place. Note that the molecule now proceeds from the amino terminal of the light chain to the carboxy terminal strand of the heavy chain. Note also that the antigen-binding region is to the left, on the other side of the molecule from the linker.

The choice of end points in TRY59 is very similar to TRY61. Linkers of this length are rare. The tension between wanting short linkers that fit very well and which could be found for the two-linker case (TRY40) and the desire to have only one linker, (which is more likely to fold correctly) is evident in the acceptance of TRY59. The linker runs from the point marked "A" in Figure 32B to the point marked "J." After five residues, the linker becomes helical. At the point marked "x," however, the side-chain of an ILE residue collides with part of the light chain. Accordingly, that residue was converted to GLY in the actual construction.

B. Genetic Constructs

The sequence of TRY59 (the single linker construction) is Met-[L-chain 1-105]-Lys-Glu-Ser-Gly-Ser-Val-Ser-Ser-Glu-Gln-Leu-Ala-Gln-Phe-Arg-Ser-Leu-Asp-[H-

chain 2-123]. The nucleotide sequence coding for this amino acid sequence and its translation is shown in Figure 28. The BglI to HindIII fragment (read clockwise) from plasmid pGX3773 containing the V_L sequence and the ClaI to BglI fragment (clockwise) from pGX3772 has been ligated with two oligonucleotides which form a fragment containing the linker sequence for TRY59 and have ClaI and HindIII ends. The ClaI and HindIII junctions within this plasmid are corrected by two successive site directed mutageneses to yield the correct construction. The ClaI to HindIII fragment from this plasmid is inserted into the O_L/P_R expression vector as in Examples 1 and 2. The resulting plasmid, pGX4908 (Figure 29) is transformed into an *E. coli* host. This strain is induced to produce the protein coded by the sequence in Figure 28 (TRY59).

Example 4

A. Computer Design

In this design an alternative method of choosing a linker to connect the light and heavy variable regions was used. A helical segment from human hemoglobin was chosen to span the major distance between the carboxy terminus of the variable light chain and the amino terminus of the variable heavy chain. This alpha helix from human hemoglobin was positioned at the rear of the F_v model using the computer graphics system. Care was taken to position the helix with its ends near the respective amino and carboxyl termini of the heavy and light chains. Care was also taken to place hydrophobic side chains in toward the F_v and hydro-

philic side chains toward the solvent. The connections between the ends of the variable regions and the hemoglobin helix were selected by the previously described computer method (EXAMPLE 1-3).

B. Genetic Constructs

The sequence of TRY104b (a single linker construction) is Met-[L-chain 1-106]-Ala-Glu-Gly-Thr-[(Hemoglobin helix)Leu-Ser-Pro-Ala-Asp-Lys-Thr-Asn-Val-Lys-Ala-Ala-Trp-Gly-Lys-Val-]Met-Thr-[H-chain 3-123]. The nucleotide sequence coding for this amino acid sequence and its translation is shown in Figure 33. The BglI to HindIII fragment (read clockwise) from plasmid pGX3773 containing the V_L sequence and the ClaI to BglI fragment (clockwise) from pGX3772 has been ligated with two oligonucleotides which form a fragment containing the linker sequence for TRY104b and have ClaI and HindIII ends. The ClaI and HindIII junctions within this plasmid are corrected by two successive site directed mutageneses to yield the correct construction. The ClaI to HindIII fragment from this plasmid is inserted into the O_L/P_R expression vector as in Examples 1-3. The resulting plasmid, pGX4910 (Figure 34) is transformed into an E. coli host. This strain is induced to produce the protein coded by the sequence in Figure 33 (TRY104b).

EXAMPLE 5

Purification of the Proteins

The single-chain antigen binding proteins from TRY40, TRY61, TRY59 and TRY104b are insoluble, and cells induced to produce these proteins show refractile bodies called inclusions upon microscopic exami-

nation. Induced cells were collected by centrifugation. The wet pellet was frozen on dry ice, then stored at -20°C. The frozen pellet was suspended in a buffer and washed in the same buffer, and subsequently the cells were suspended in the same buffer. The cells were broken by passage through a French pressure cell, and the inclusion bodies containing the single-chain antigen binding protein (SCA) were purified by repeated centrifugation and washing. The pellet was solubilized in guanidine-HCl, and reduced with 2-mercaptoethanol. The solubilized material was passed through a gel filtration column, i.e., Sephacryl™ S-300. Other methods such as ion exchange could be used.

EXAMPLE 6

Folding of the Proteins

Purified material was dialyzed against water, and the precipitate protein collected by centrifugation. The protein was solubilized in urea and reduced with 2-mercaptoethanol. This denatured and solubilized material was dialyzed against a buffer containing salt and reducing agents to establish the redox potential to form the intra domain (one each for the light and heavy chain variable region sequences) disulfide bridges (Saxena and Wetlanfer, Biochem 9:5015-5023 (1970)). The folded protein was assayed for BGH binding activity.

The TRY59 protein used in competition experiments was solubilized and renatured directly from inclusions. This material was subsequently purified by affinity to BGH-Sepharose.

EXAMPLE 7

Binding Assay

BGH was immobilized on nitrocellulose strips along with non-specific proteins such as bovine serum albumin or lysozymes. Further non-specific protein binding was blocked with an immunologically inert protein, for example gelatin. Folded SCA was tested for its ability to bind to BGH. The SCA was detected by a rabbit anti-L chain (of the monoclonal) anti-serum. The rabbit antibodies were reacted with goat anti-rabbit IgG coupled to peroxidase. The strips were reacted with chemicals which react with the peroxidase to give a color reaction if the peroxidase is present.

Figure 35 shows the result of this spot assay for TRY61 (strip 1) and TRY40 (strip 2). Strip 3 was stained with amido black to show the presence of all three proteins. The other proteins, TRY59, TRY104b gave similar results in the spot assay. A competition assay with the SCA competing with the monoclonal can be used as well. The results of competing the F_{ab} of 3C2 monoclonal with 1 and 10 ug of TRY59 protein which had been affinity purified are shown in Figure 36 (▲ F_{ab} alone, ■ F_{ab} + 1 ug TRY59, and ● F_{ab} + 10 ug TRY59). The affinity estimated from the Ic_{50} of this experiment was approximately 10^6. The data are summarized in Table 1.

Table 1

Properties of the Proteins Produced by the Single-chain Constructions

Single-chain	Number of Linkers	Spot Test	Binding to BGH Column	Ka
TRY40	2	+	–	ND
TRY59	1	+	+	~10^{8}
TRY61	1	+	–	ND
TRY104B	1	+	+	ND
3C2 monoclonal	NA	+	+	10^{8} – 10

ND = not determined
NA = not applicable

WE CLAIM:

1. A single polypeptide chain binding molecule which has binding specificity substantially similar to the binding specificity of the light and heavy chain aggregate variable region of an antibody.

2. The molecule of claim 1 which comprises two peptide linkers joining said light and heavy chains into said single chain.

3. The molecule of claim 2 which comprises in sequence:

(a) an N-terminal region derived from said light chain;
(b) a peptide linker;
(c) a peptide region derived from said heavy chain;
(d) a second peptide linker; and
(e) a C-terminal region derived from said light chain.

4. The molecule of claim 1 which comprises one peptide linker joining said light and heavy chains into said single chain.

5. The molecule of claim 4 which comprises, in sequence:

(a) an N-terminal region derived from said light chain;
(b) a peptide linker; and
(c) a C-terminal region derived from said heavy chain.

6. The molecule of claim 4 which comprises in sequence:
   (a) an N-terminal region derived from said heavy chain;
   (b) a peptide linker; and
   (c) a C-terminal region derived from said light chain.

7. The molecule of claim 3, 5 or 6 which, prior to said N-terminal region (a), comprises a methionine residue.

8. The molecule of claim 1 which is detectably labeled.

9. The molecule of claim 1 which is in immobilized form.

10. The molecule of claim 1 which is conjugated to an imaging agent.

11. The molecule of claim 1 which is conjugated to a toxin.

12. A genetic sequence coding for the molecule of claim 1.

13. A recombinant DNA (rDNA) molecule comprising the sequence of claim 12.

14. The rDNA molecule of claim 13 which is a replicable cloning or expression vehicle.

15. The rDNA molecule of claim 14 wherein said vehicle is a plasmid.

16. A host cell transformed with the rDNA molecule of claim 13.

17. The host cell of claim 16 which is a bacterial cell, a yeast or other fungal cell or a mammalian cell line in vitro.

18. A method of producing a single polypeptide chain binding molecule which has binding specificity substantially similar to the binding specificity of the light and heavy chain aggregate variable region of an antibody, which comprises:

(a) providing a genetic sequence coding for said molecule;
(b) transforming a host cell with said sequence;
(c) expressing said sequence in said host; and
(d) recovering said molecule.

19. The method of claim 18 which further comprises purifying said recovered molecule.

20. The method of claim 18 wherein said host cell is a bacterial cell, yeast or other fungal cell, or a mammalian cell line.

21. The binding molecule produced by the method of claim 18 or 19.

22. In an immunoassay method which utilizes an antibody in labeled form, the improvement comprising using the molecule of claim 8 instead of said antibody.

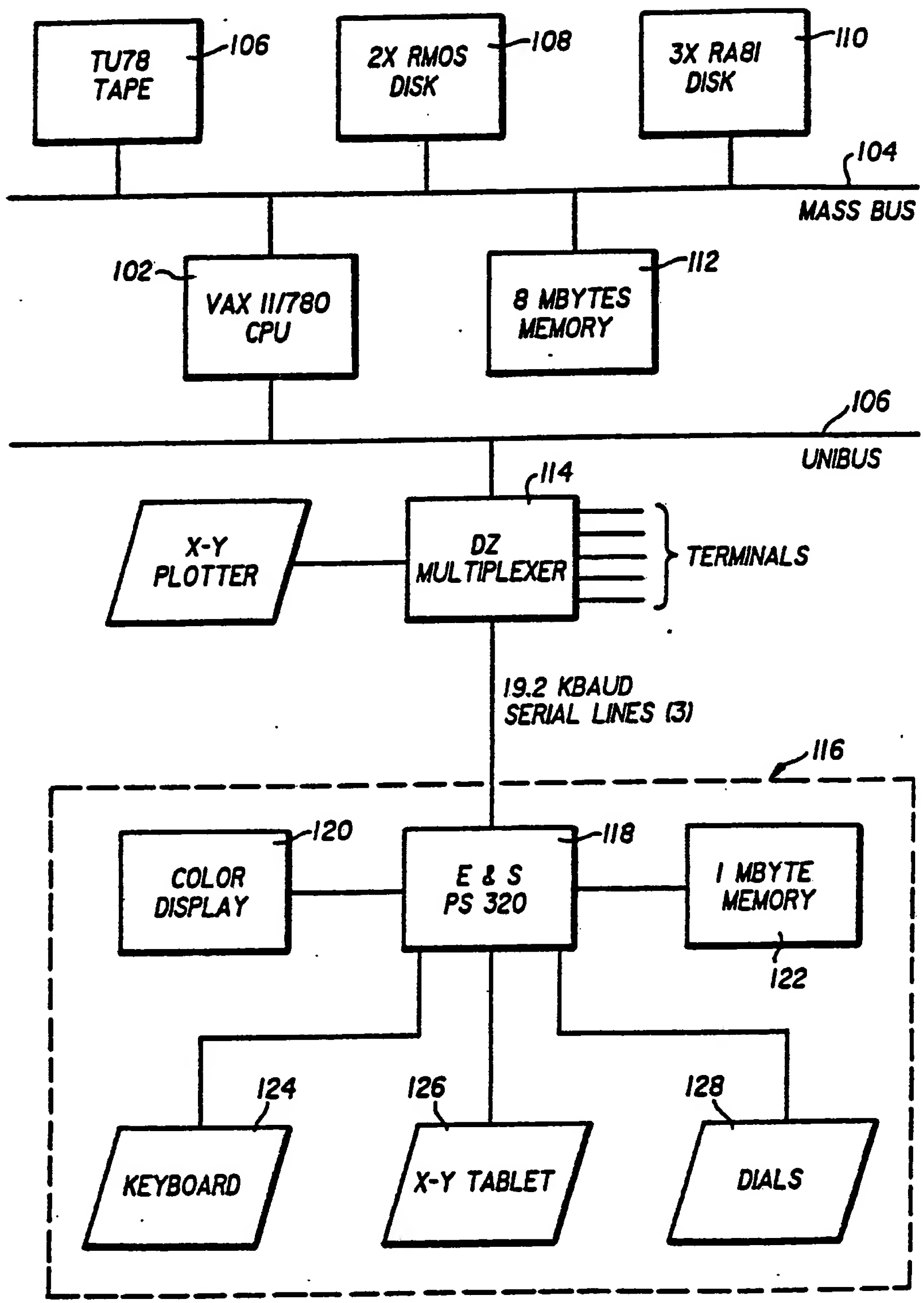
23. In an immunoassay method which utilizes an antibody in immobilized form, the improvement comprising using the molecule of claim 9 instead of said antibody.

24. In the immunoassay of claim 21 or 22 wherein said immunoassay is a competitive immunoassay.

25. In the immunoassay of claim 21 or 22 wherein said immunoassay is a sandwich immunoassay.

26. In an immunotherapeutic method which utilizes an antibody conjugated to a therapeutic agent, the improvement comprising using the molecule of claim 1 instead of said antibody.

27. In a method of immunoaffinity purification which utilizes an antibody therefor, the improvement which comprises using the molecule of claim 1 instead of said antibody.



TU78 TAPE
106
2X RM05 DISK
108
3X RA81 DISK
110
104
MASS BUS
102
VAX 11/780 CPU
8 MBYTES MEMORY
112
106
UNIBUS
114
X-Y PLOTTER
DZ MULTIPLEXER
TERMINALS
19.2 KBAUD SERIAL LINES (3)
116
120
COLOR DISPLAY
118
E & S PS 320
1 MBYTE MEMORY
122
124
126
128
KEYBOARD
X-Y TABLET
DIALS

FIG. 1

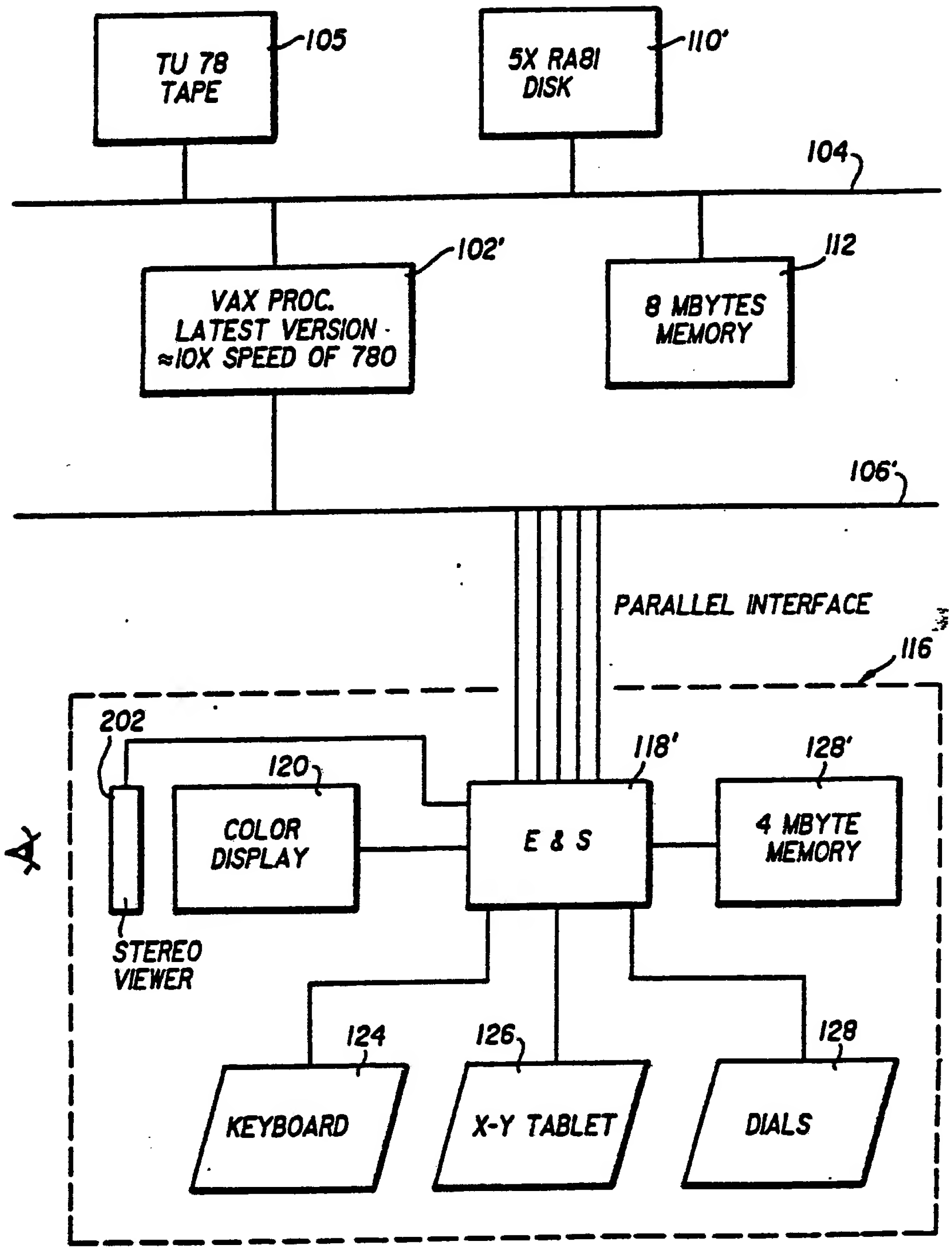

TU 78
TAPE
105
5X RA81
DISK
110'
104
VAX PROC.
LATEST VERSION
≈10X SPEED OF 780
102'
8 MBYTES
MEMORY
112
106'
PARALLEL INTERFACE
116
202
120
118'
128'
COLOR
DISPLAY
E & S
4 MBYTE
MEMORY
STEREO
VIEWER
124
126
128
KEYBOARD
X-Y TABLET
DIALS

FIG. 2

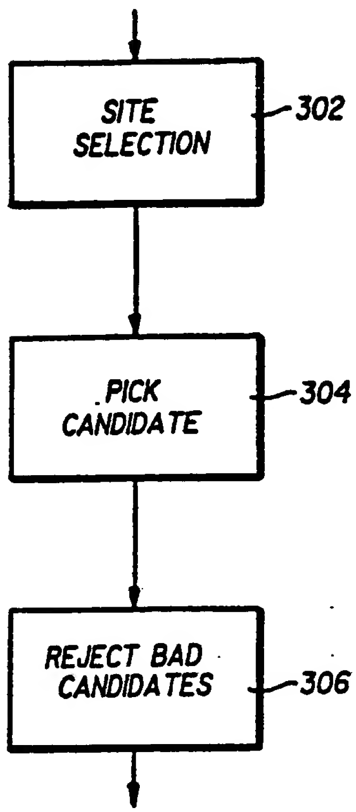


FIG.3

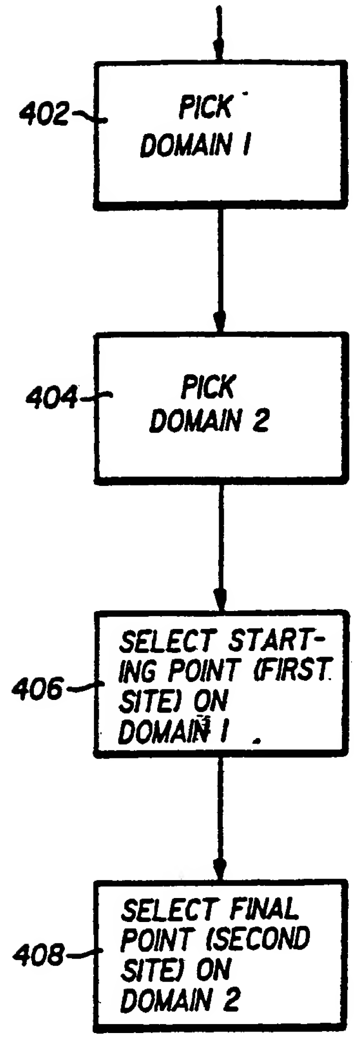


FIG.4

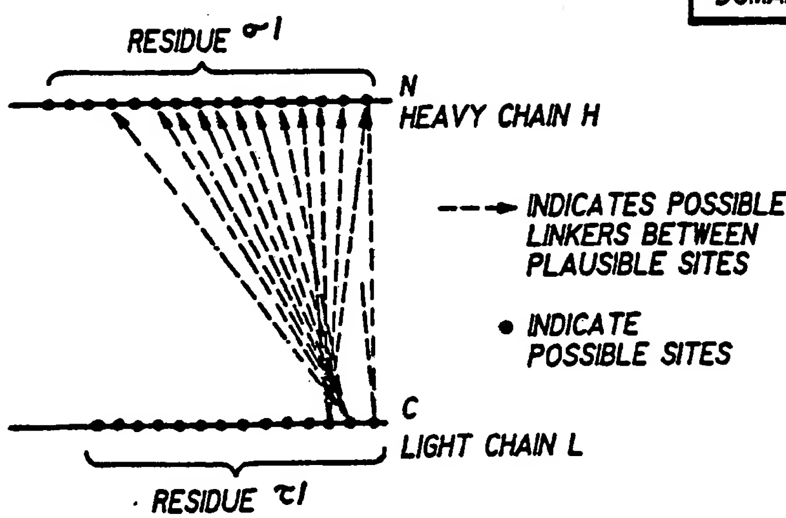


FIG.7

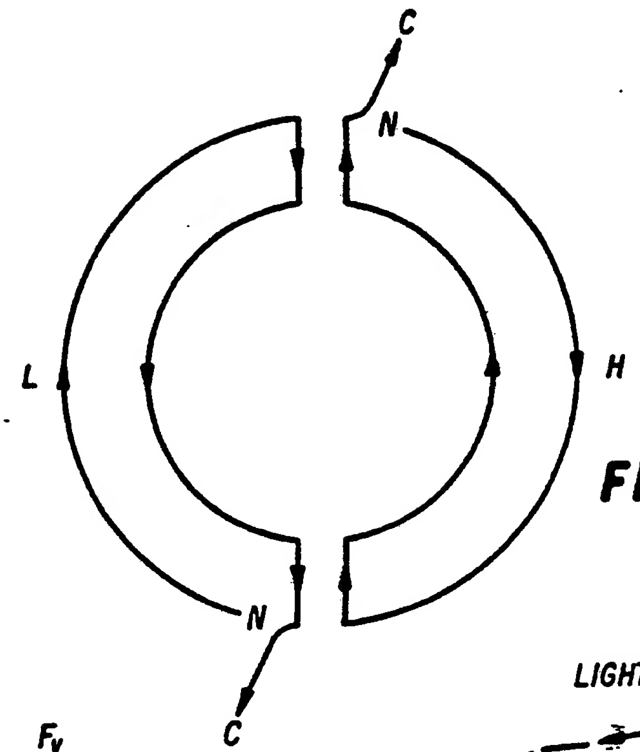


FIG. 5A

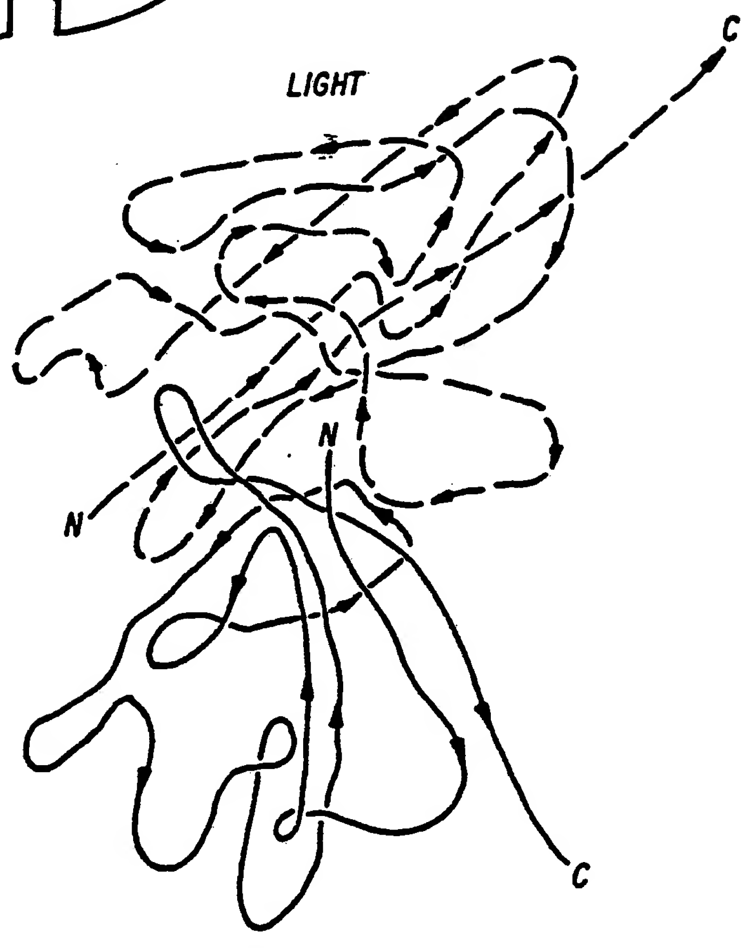


FIG. 5B

SUBSTITU...

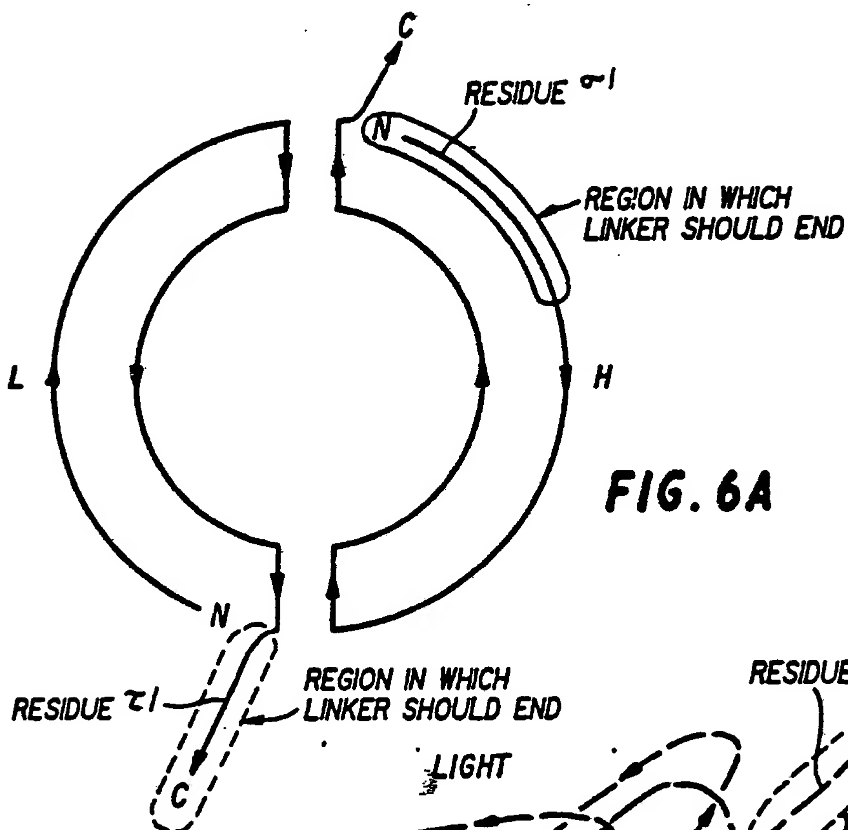

C
RESIDUE σ1
N
REGION IN WHICH
LINKER SHOULD END
L
H
N
RESIDUE τ1
REGION IN WHICH
LINKER SHOULD END
C

FIG. 6A

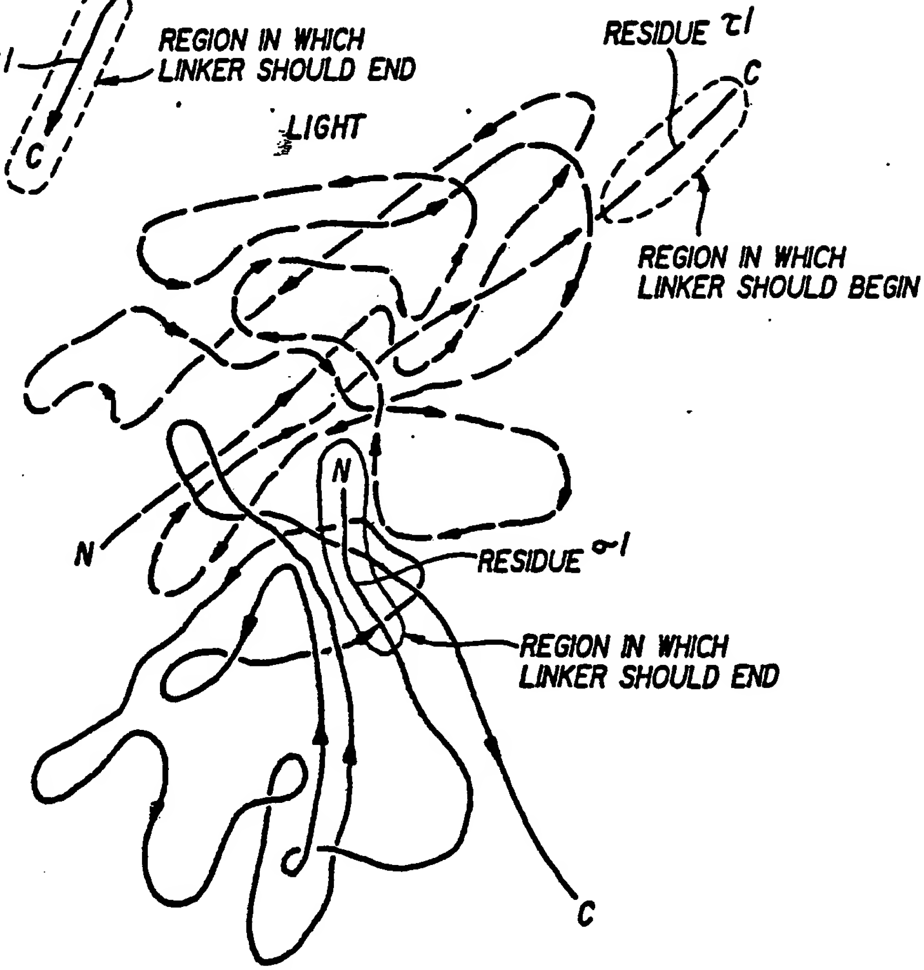

LIGHT
RESIDUE τ1
C
REGION IN WHICH
LINKER SHOULD BEGIN
N
N
RESIDUE σ1
REGION IN WHICH
LINKER SHOULD END
C
HEAVY

FIG. 6B

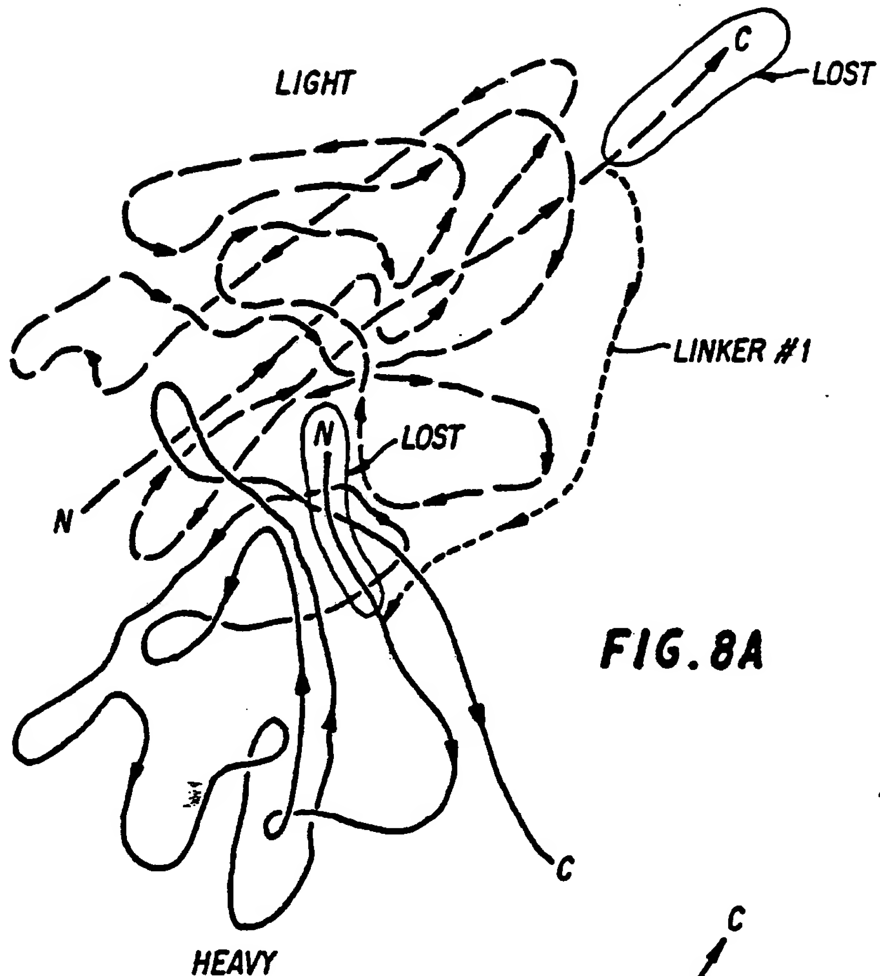

LIGHT
C
LOST
LINKER #1
N
LOST
N
C
C
HEAVY

FIG. 8A

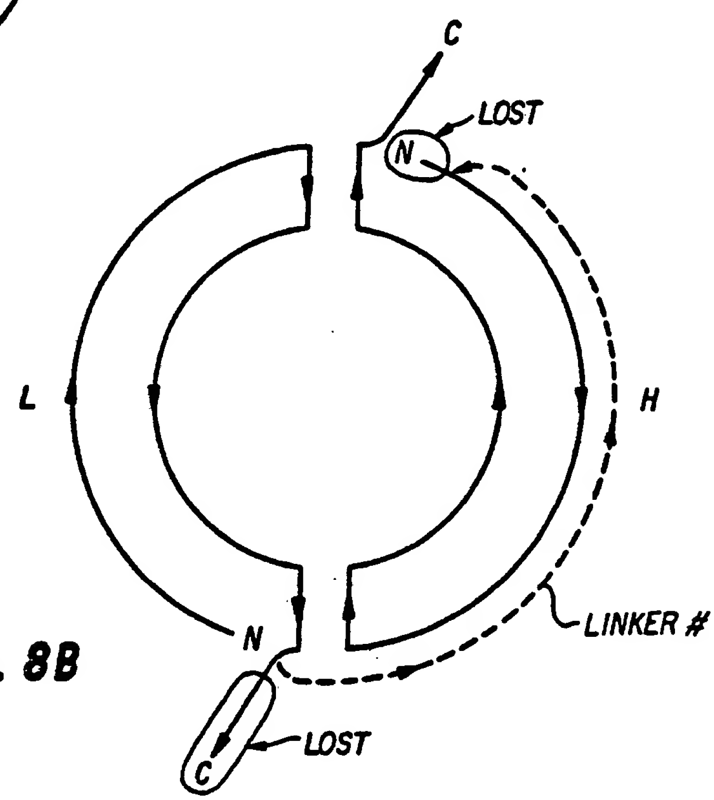

C
LOST
N
L
H
N
LINKER #1
C
LOST

FIG. 8B

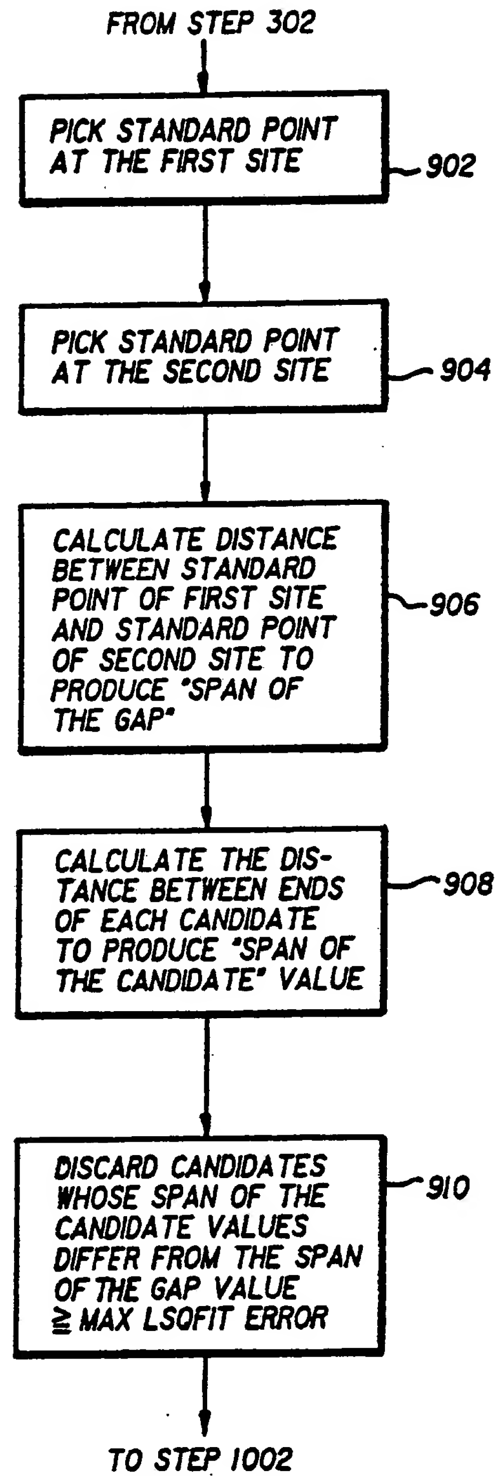

FROM STEP 302
PICK STANDARD POINT AT THE FIRST SITE
902
PICK STANDARD POINT AT THE SECOND SITE
904
CALCULATE DISTANCE BETWEEN STANDARD POINT OF FIRST SITE AND STANDARD POINT OF SECOND SITE TO PRODUCE "SPAN OF THE GAP"
906
CALCULATE THE DISTANCE BETWEEN ENDS OF EACH CANDIDATE TO PRODUCE "SPAN OF THE CANDIDATE" VALUE
908
DISCARD CANDIDATES WHOSE SPAN OF THE CANDIDATE VALUES DIFFER FROM THE SPAN OF THE GAP VALUE ≧ MAX LSQFIT ERROR
910
TO STEP 1002

FIG. 9

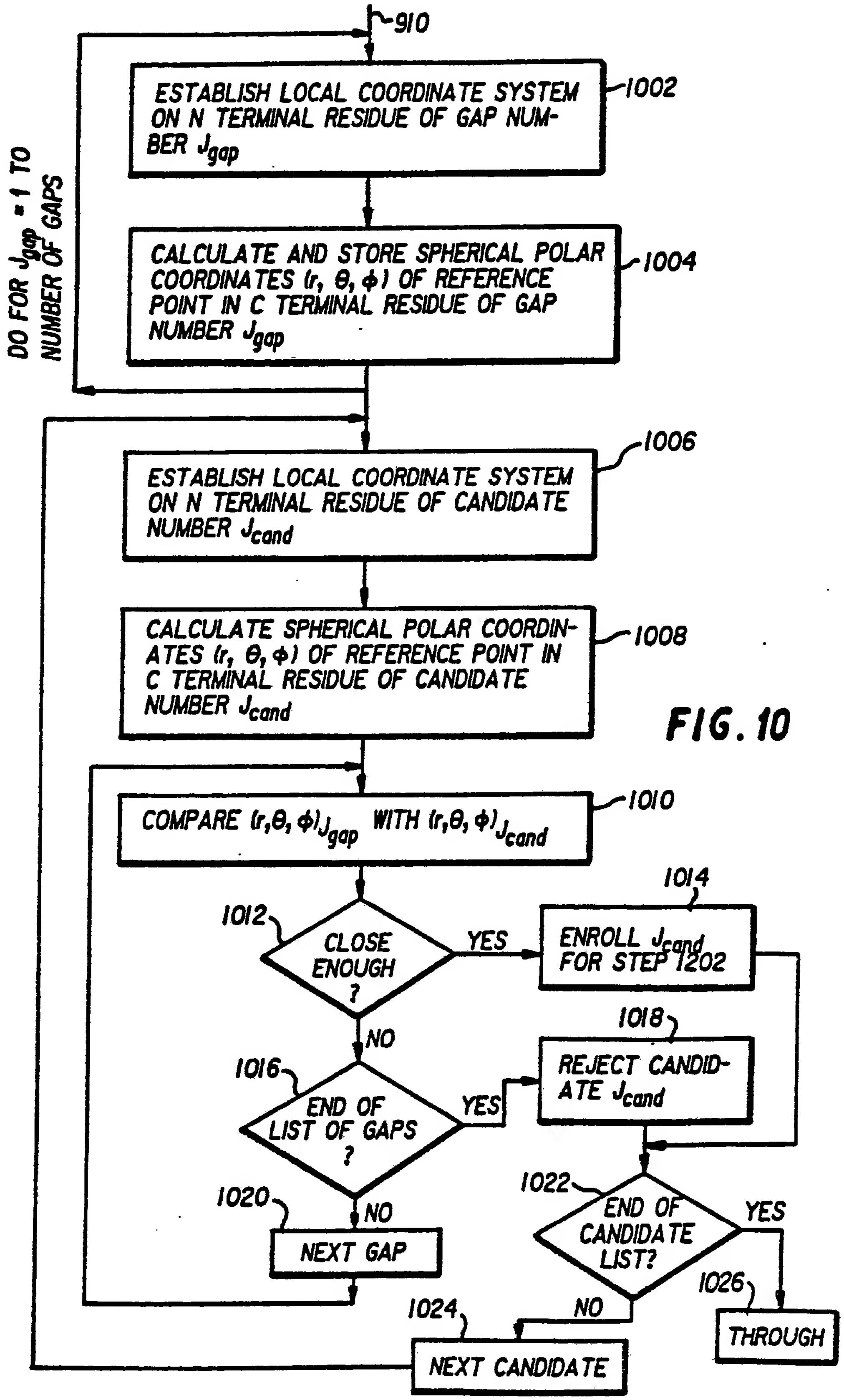

910
ESTABLISH LOCAL COORDINATE SYSTEM ON N TERMINAL RESIDUE OF GAP NUMBER J_{gap}
1002
CALCULATE AND STORE SPHERICAL POLAR COORDINATES (r, θ, ϕ) OF REFERENCE POINT IN C TERMINAL RESIDUE OF GAP NUMBER J_{gap}
1004
DO FOR J_{gap} = 1 TO NUMBER OF GAPS
ESTABLISH LOCAL COORDINATE SYSTEM ON N TERMINAL RESIDUE OF CANDIDATE NUMBER J_{cand}
1006
CALCULATE SPHERICAL POLAR COORDINATES (r, θ, ϕ) OF REFERENCE POINT IN C TERMINAL RESIDUE OF CANDIDATE NUMBER J_{cand}
1008
COMPARE $(r,\theta,\phi)_{J_{gap}}$ WITH $(r,\theta,\phi)_{J_{cand}}$
1010
1012
CLOSE ENOUGH ?
YES
1014
ENROLL J_{cand} FOR STEP 1202
NO
1016
END OF LIST OF GAPS ?
YES
1018
REJECT CANDIDATE J_{cand}
1020
NO
NEXT GAP
1022
END OF CANDIDATE LIST?
YES
NO
1024
NEXT CANDIDATE
1026
THROUGH

FIG. 10

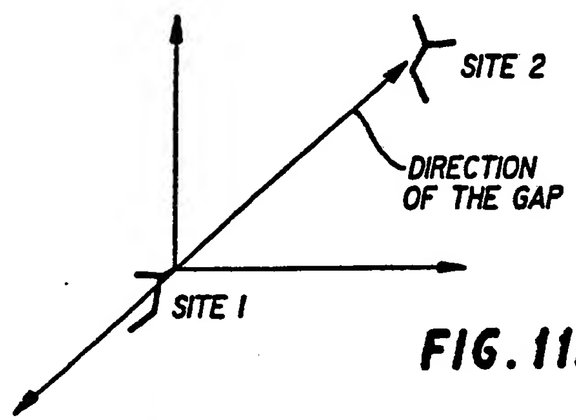


FIG. 11A

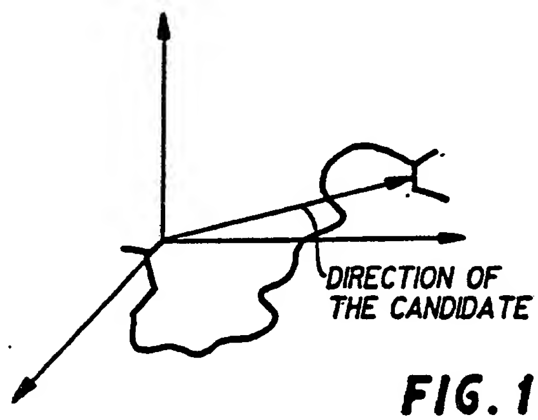


FIG. 11B

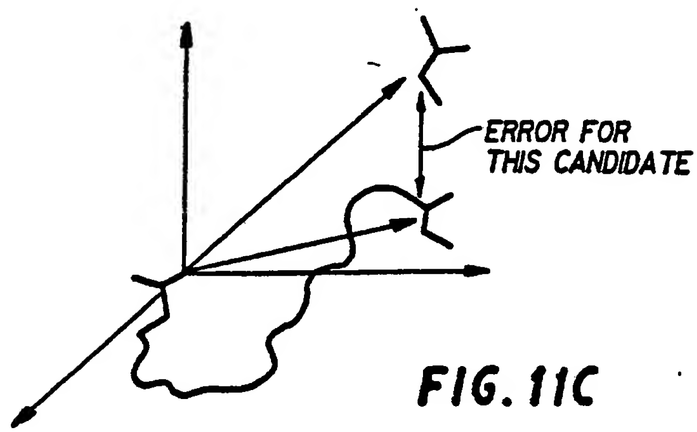


FIG. 11C

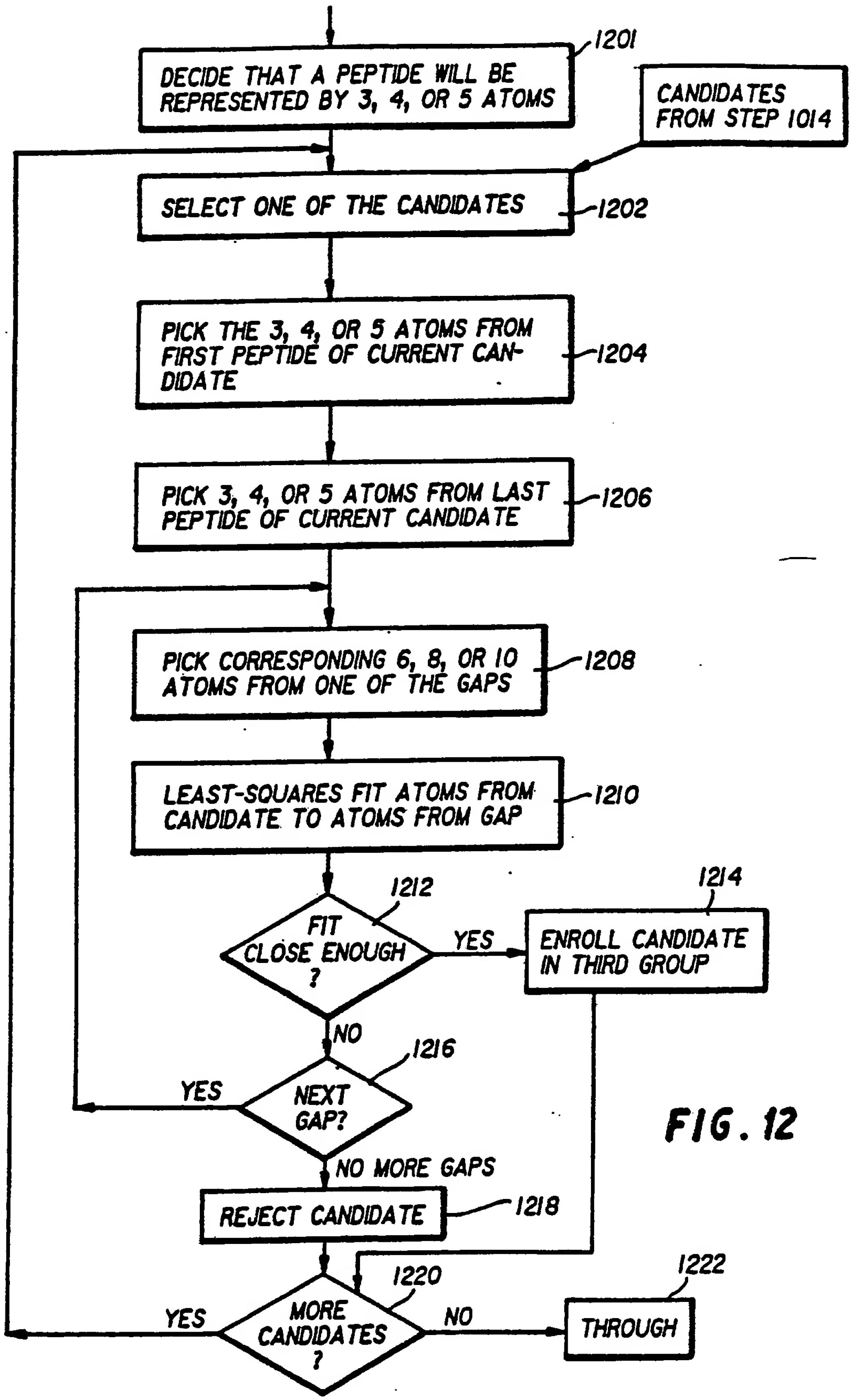

DECIDE THAT A PEPTIDE WILL BE REPRESENTED BY 3, 4, OR 5 ATOMS
1201
CANDIDATES FROM STEP 1014
SELECT ONE OF THE CANDIDATES
1202
PICK THE 3, 4, OR 5 ATOMS FROM FIRST PEPTIDE OF CURRENT CANDIDATE
1204
PICK 3, 4, OR 5 ATOMS FROM LAST PEPTIDE OF CURRENT CANDIDATE
1206
PICK CORRESPONDING 6, 8, OR 10 ATOMS FROM ONE OF THE GAPS
1208
LEAST-SQUARES FIT ATOMS FROM CANDIDATE TO ATOMS FROM GAP
1210
FIT CLOSE ENOUGH ?
1212
YES
ENROLL CANDIDATE IN THIRD GROUP
1214
NO
NEXT GAP?
1216
YES
NO MORE GAPS
REJECT CANDIDATE
1218
MORE CANDIDATES ?
1220
YES
NO
THROUGH
1222

FIG. 12

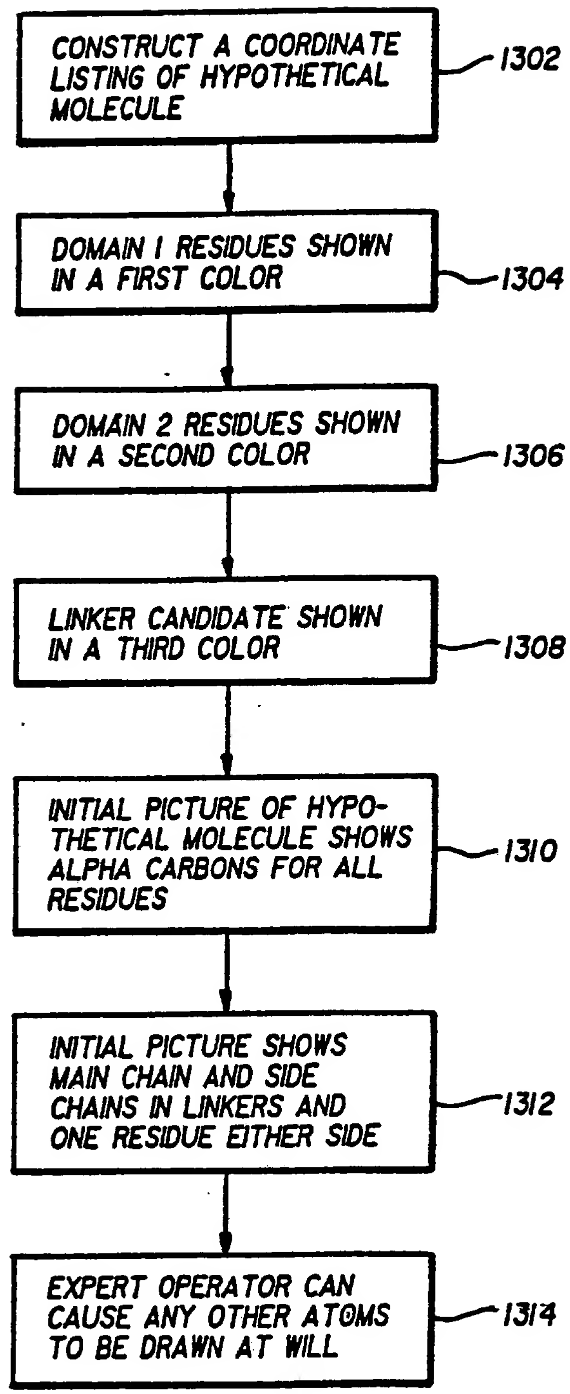

CONSTRUCT A COORDINATE LISTING OF HYPOTHETICAL MOLECULE
1302
DOMAIN 1 RESIDUES SHOWN IN A FIRST COLOR
1304
DOMAIN 2 RESIDUES SHOWN IN A SECOND COLOR
1306
LINKER CANDIDATE SHOWN IN A THIRD COLOR
1308
INITIAL PICTURE OF HYPOTHETICAL MOLECULE SHOWS ALPHA CARBONS FOR ALL RESIDUES
1310
INITIAL PICTURE SHOWS MAIN CHAIN AND SIDE CHAINS IN LINKERS AND ONE RESIDUE EITHER SIDE
1312
EXPERT OPERATOR CAN CAUSE ANY OTHER ATOMS TO BE DRAWN AT WILL
1314

FIG. 13

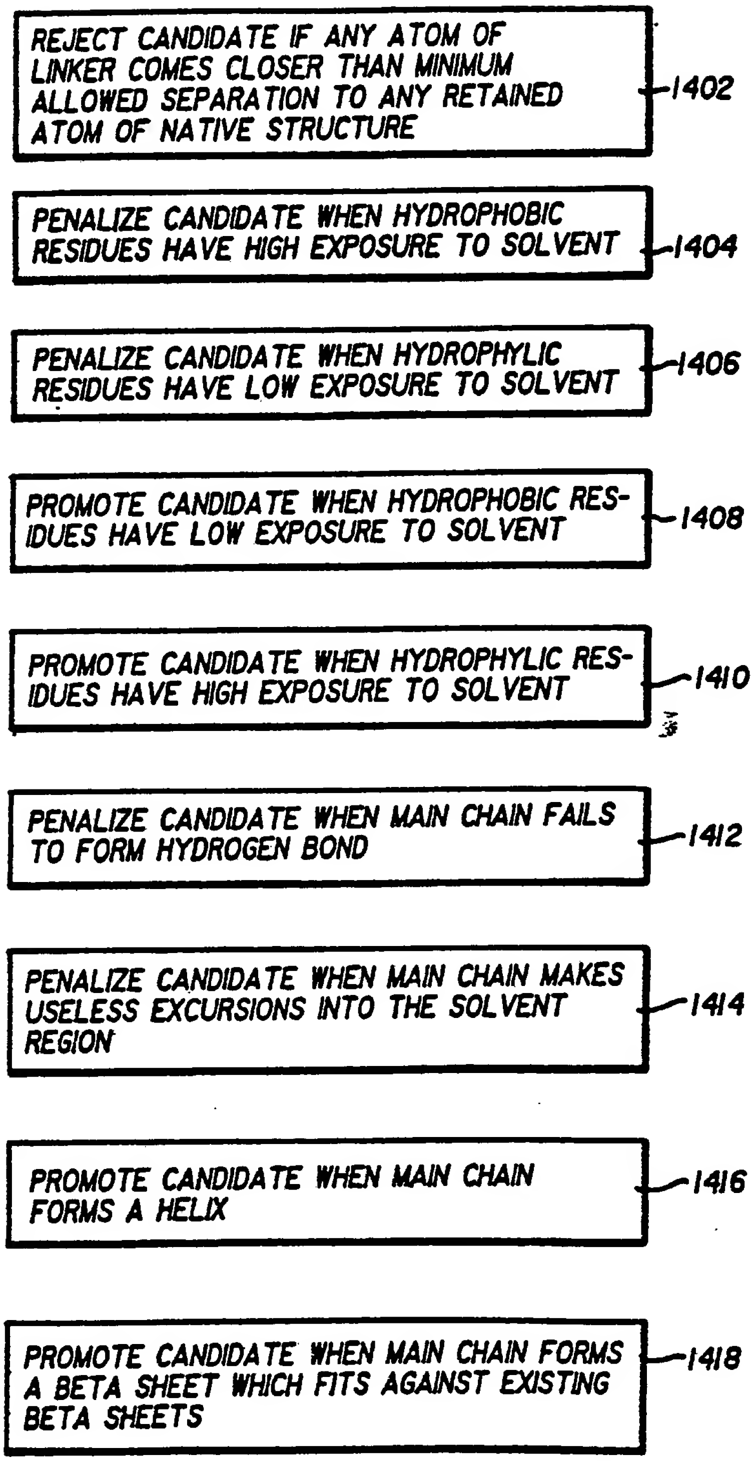

REJECT CANDIDATE IF ANY ATOM OF LINKER COMES CLOSER THAN MINIMUM ALLOWED SEPARATION TO ANY RETAINED ATOM OF NATIVE STRUCTURE
1402
PENALIZE CANDIDATE WHEN HYDROPHOBIC RESIDUES HAVE HIGH EXPOSURE TO SOLVENT
1404
PENALIZE CANDIDATE WHEN HYDROPHYLIC RESIDUES HAVE LOW EXPOSURE TO SOLVENT
1406
PROMOTE CANDIDATE WHEN HYDROPHOBIC RESIDUES HAVE LOW EXPOSURE TO SOLVENT
1408
PROMOTE CANDIDATE WHEN HYDROPHYLIC RESIDUES HAVE HIGH EXPOSURE TO SOLVENT
1410
PENALIZE CANDIDATE WHEN MAIN CHAIN FAILS TO FORM HYDROGEN BOND
1412
PENALIZE CANDIDATE WHEN MAIN CHAIN MAKES USELESS EXCURSIONS INTO THE SOLVENT REGION
1414
PROMOTE CANDIDATE WHEN MAIN CHAIN FORMS A HELIX
1416
PROMOTE CANDIDATE WHEN MAIN CHAIN FORMS A BETA SHEET WHICH FITS AGAINST EXISTING BETA SHEETS
1418

FIG. 14

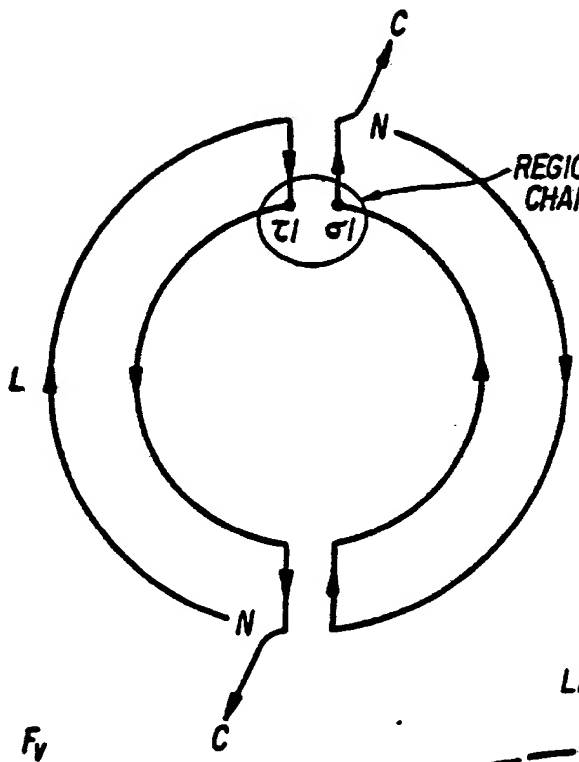


FIG. 15A

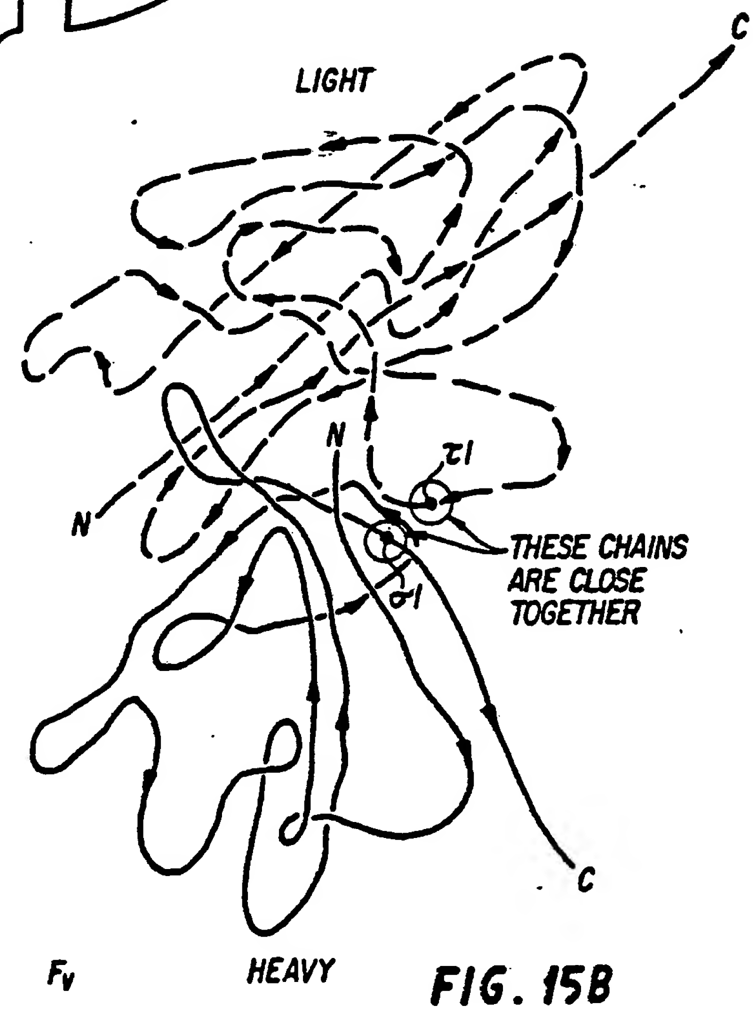


FIG. 15B

SUBSTITUTE SHEET

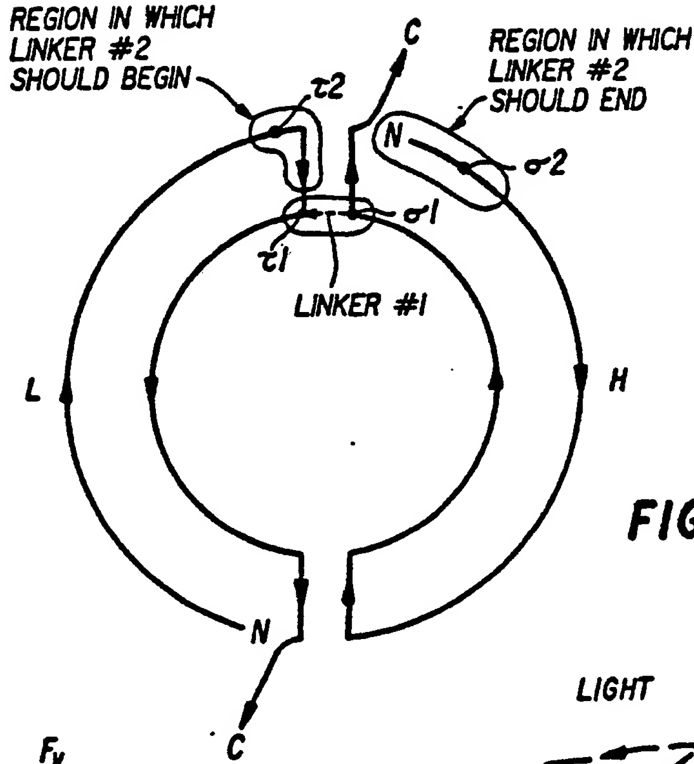


FIG. 16A

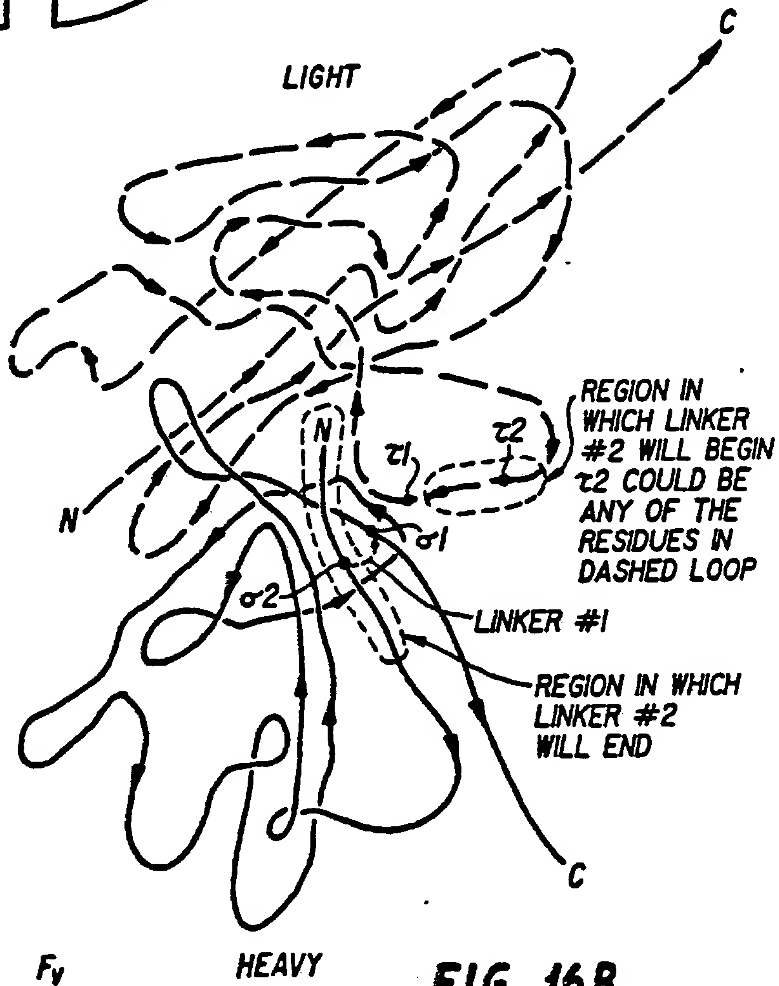


FIG. 16B

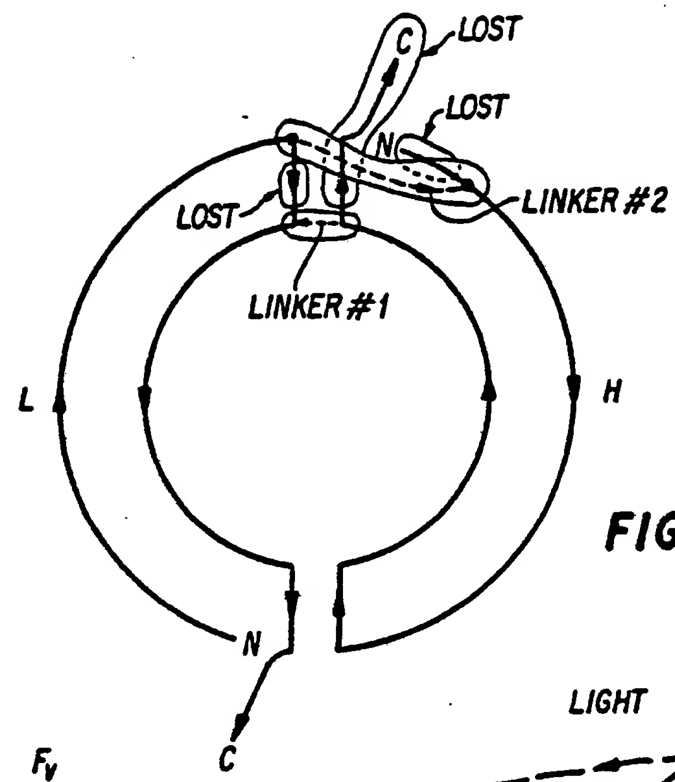


FIG. 17A

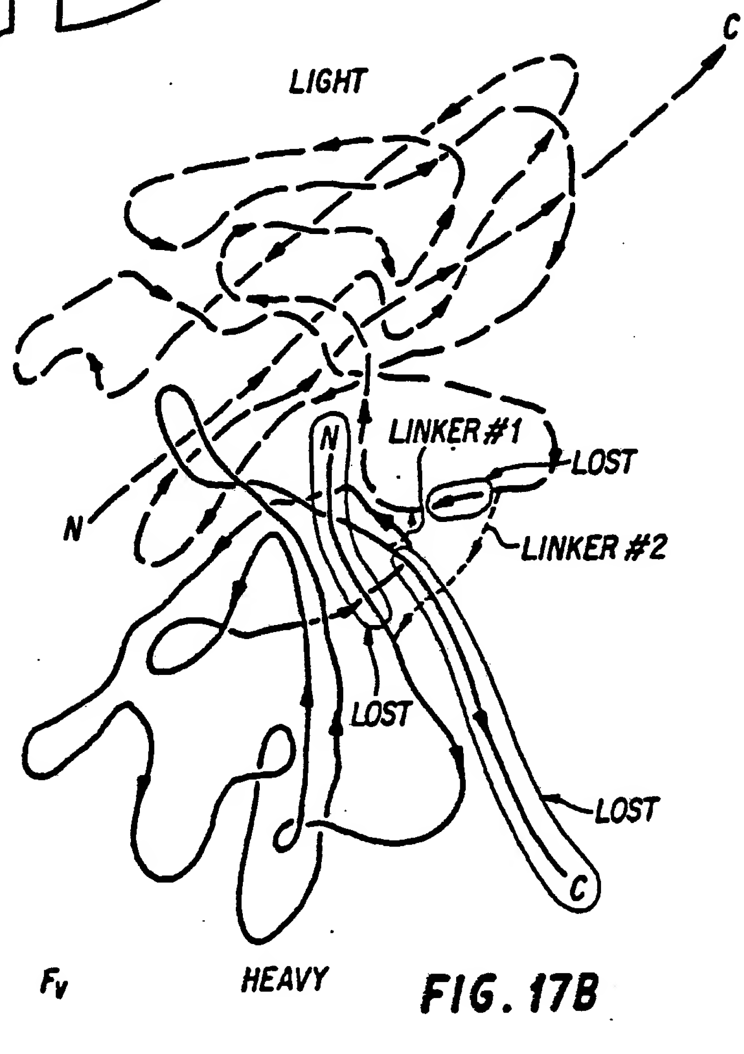


FIG. 17B

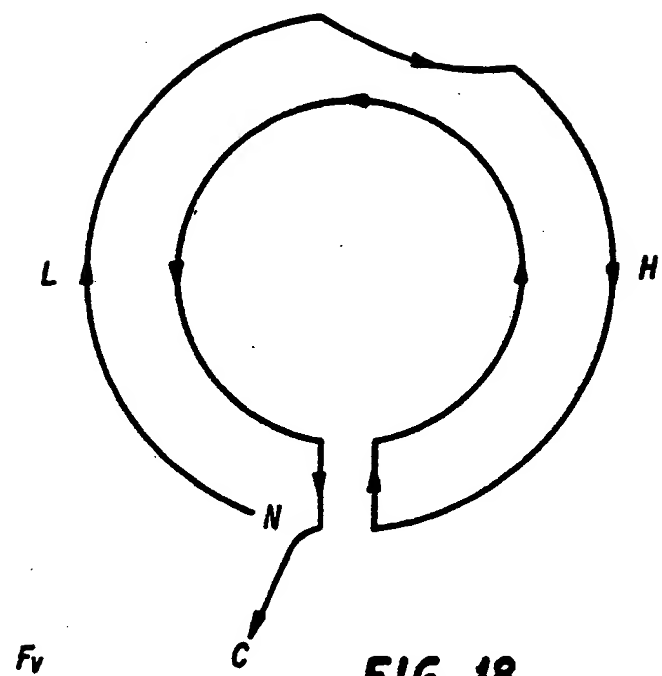


FIG. 18

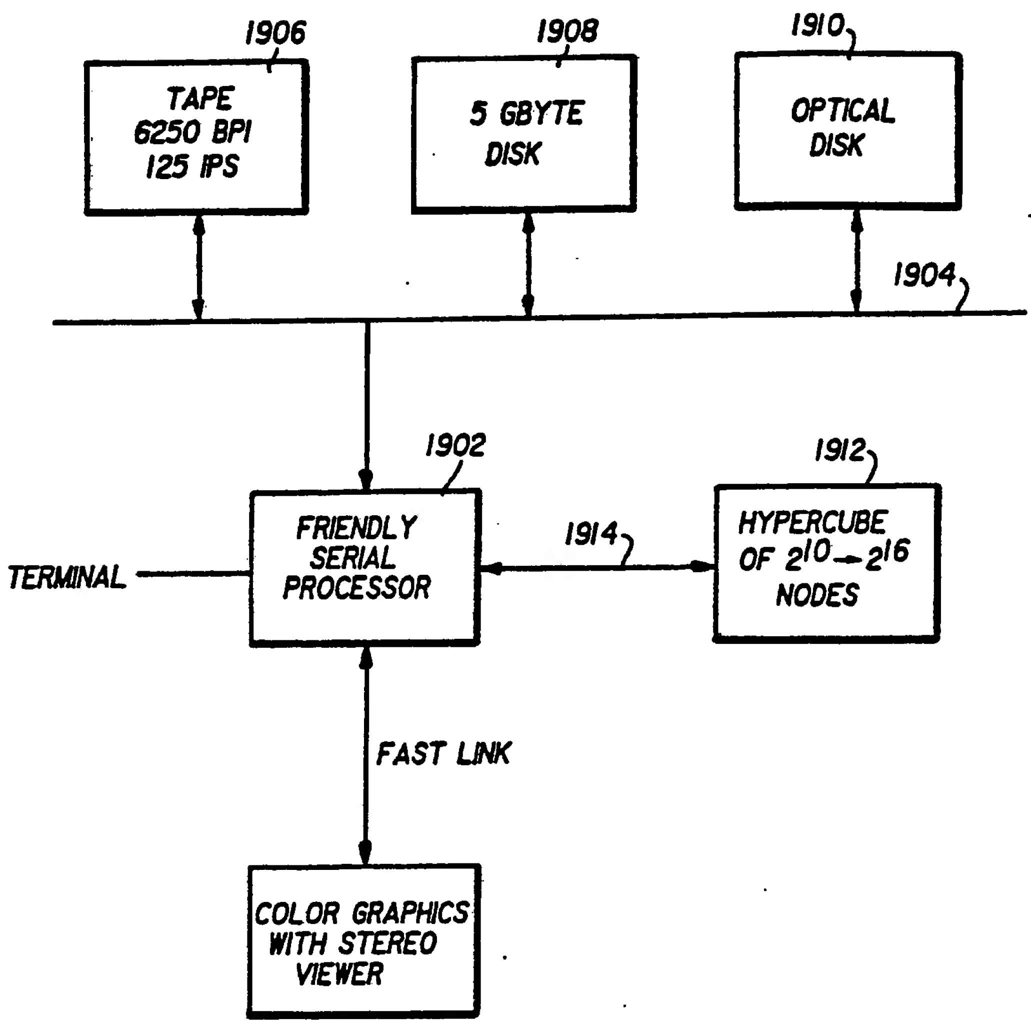

1906
TAPE
6250 BPI
125 IPS
1908
5 GBYTE
DISK
1910
OPTICAL
DISK
1904
1902
FRIENDLY
SERIAL
PROCESSOR
TERMINAL
1914
1912
HYPERCUBE
OF $2^{10} \rightarrow 2^{16}$
NODES
FAST LINK
COLOR GRAPHICS
WITH STEREO
VIEWER

FIG. 19

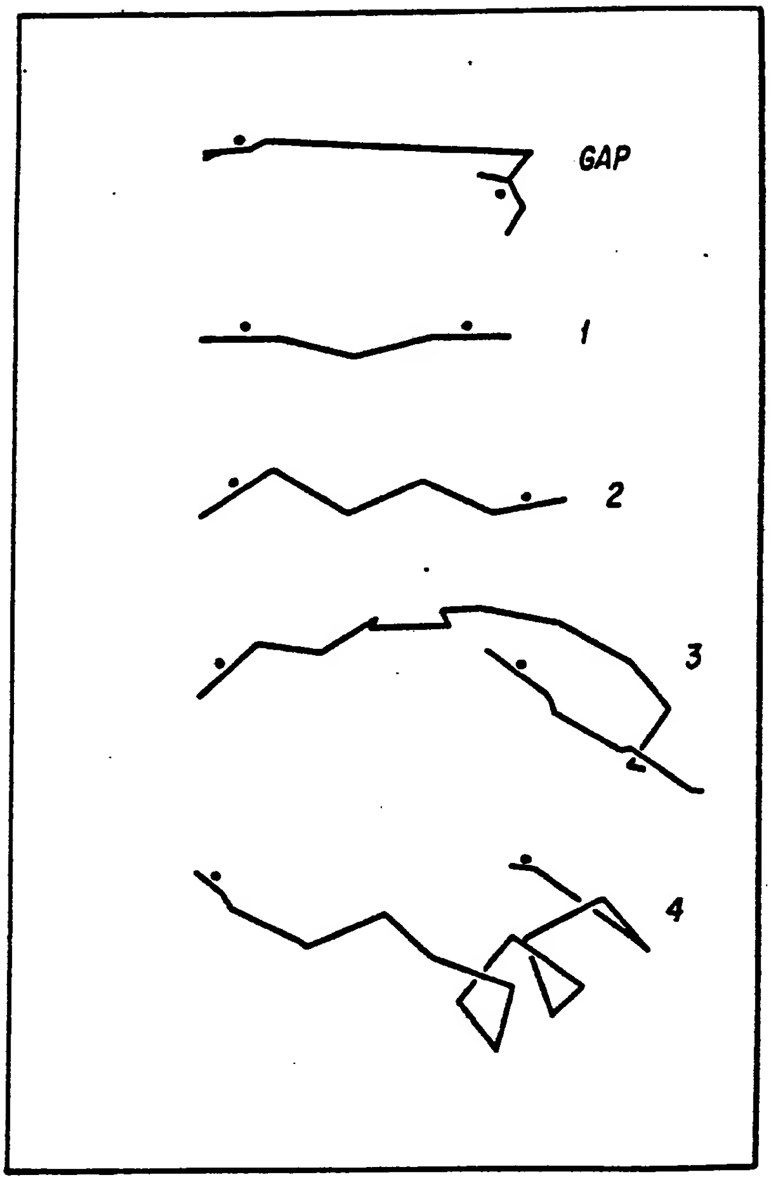


FIG. 20A

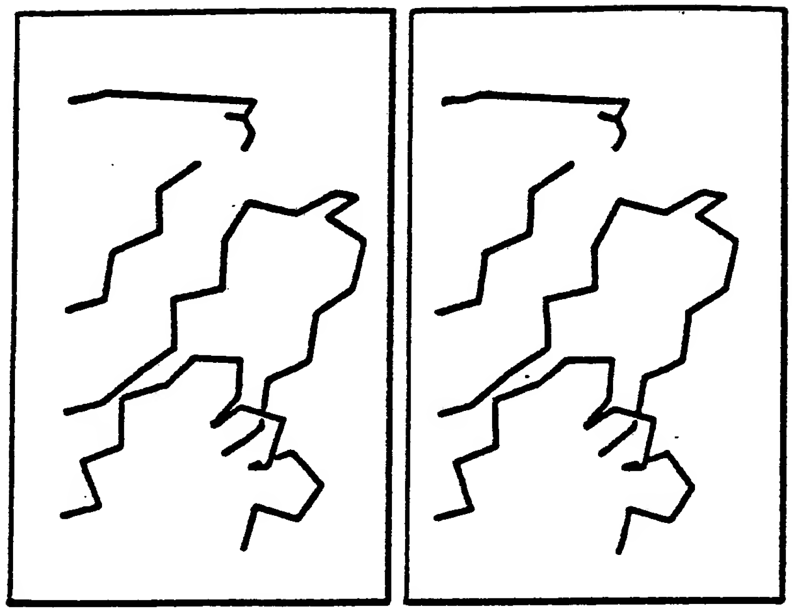


FIG. 20B

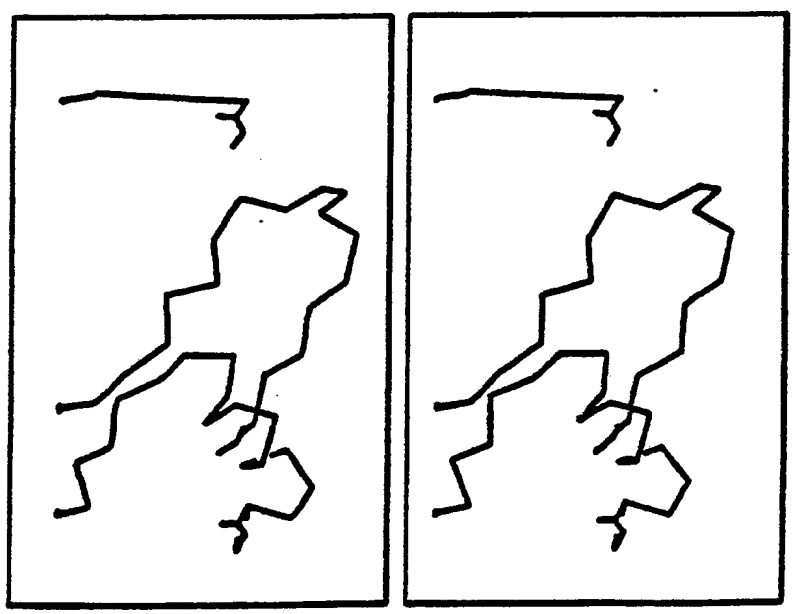


FIG. 20C

5 10 15 20
Glu Val Gln Leu Val Glu Ser Gly Gly Asp Leu Val Lys Pro Gly Gly Ser Leu Lys Leu
GAG GTG CAC CTG GTG GAG TCT GGG GGA GAC TTA GTG AAG CCT GGA GGG TCC CTG AAA CTC

25 30 35 40
Ser Cys Ala Ala Ser Gly Phe Thr Phe Ile Ser Tyr Gly Met Ser Trp Val Arg Gln Thr
TCC TGT GCA GCC TCT GGA TTC ACT TTC ATT AGC TAT GGC ATG TCT TGG GTT CGC CAG ACT

45 50 55 60
Pro Asp Lys Arg Leu Glu Trp Val Ala Thr Ile Ser Ser Gly Ser Thr Tyr Thr Tyr Tyr
CCA GAC AAG AGG CTG GAG TGG GTC GCA ACC ATT AGT AGT GGT AGT ACT TAC ACC TAC TAT

65 70 75 80
Pro Asp Ser Val Lys Gly Arg Phe Thr Ile Ser Arg Asp Asn Ala Lys Asn Thr Leu Tyr
CCA GAC AGT GTG AAG GGG CGA TTC ACC ATC TCC AGA GAC AAT GCC AAG AAC ACC CTG TAC

85 90 95 100
Leu Gln Met Ser Gly Leu Lys Ser Glu Asp Thr Ala Met Tyr Tyr Cys Ala Arg Arg Ile
CTG CAA ATG AGC GGT CTG AAG TCT GAG GAC ACA GCC ATG TAT TAC TGT GCA AGA CGG ATT

105 110 115 120
Thr Thr Val Val Leu Thr Asp Tyr Tyr Ala Met Asp Tyr Trp Gly Gln Gly Thr Ser Val
AGT ACG GTA GTA CTT ACG GAT TAC TAT GCT ATG GAC TAC TGG GGT CAA GGA ACC TCA GTC

125 130 135 140
Thr Val Ser Ser Ala Lys Thr Thr Pro Pro Ser Val Tyr Pro leu Ala Pro Gly Ser Ala
ACC GTC TCC TCA GCC AAA ACG ACA CCC CCA TCT GTC TAT CCA CTG GCC CCT GGA TCT GCT

145 150 155 160
Ala Gln Thr Asn Ser Met Val Thr Leu Gly Cys Leu Val Lys Gly Tyr Phe Pro Glu Pro
GCC CAA ACT AAC TCG ATG GTG ACC CTG GGA TGC CTG GTC AAG GGC TAT TTC CCT GAG CCA

165 170 175 180
Val Thr Val Thr Trp Asn Ser Gly Ser Leu Ser Ser Gly Val His Thr Phe Pro Ala Val
GTG ACA GTG ACC TGG AAC TCT GGA TCC CTG TCC AGC GGT GTG CAC ACC TTC CCA GCT GTC

185 190 195 200
Leu Gln Ser Asp Leu Tyr Thr Leu Ser Ser Ser Val Thr Val Pro Ser Ser Thr Trp Pro
CTG CAG TCT GAC CTC TAC ACT CTG AGC AGC TCA GTG ACT GTG CCC TCC AGC ACC TGG CCC

205 210 215 220
Ser Glu Thr Val Thr Cys Asn Val Ala His Pro Ala Ser Ser Thr Lys Val Asp Lys Lys
AGC GAG ACC GTC ACC TGC AAC GTT GCC CAC CCG GCG AGC ACC ACC AAG GTG GAC AAG AAA

225 230 235 240
Ile Val Pro Arg Asp Cys Gly Cys Lys Pro Cys Ile Cys Thr Val Pro Glu Val Ser Ser
ATT GTG CCC AGG GAT TGT GGT TGT AAG CCT TGC ATA TGT ACA GTC CCA GAA GTA TCA TCT

245 250 255 260
Val Phe Ile Phe Pro Pro Lys Pro Lys Asp Val Leu Thr Ile Thr Leu Thr Pro Lys Val
GTC TTC ATC TTC CCC CCA AAG CCC AAG GAT GTG CTC ACC ATT ACT CTG ACT CCT AAG GTC

265 270 275 280
Thr Cys Val Val Val Asp Ile Ser Lys Asp Asp Pro Glu Val Gln Phe Ser Trp Phe Val
ACG TGT GTT GTG GTA GAC ATC AGC AAG GAT GAT CCC GAG GTC CAG TTC AGC TGG TTT GTA

285 290 295 300
Asp Asp Val Glu Val His Thr Ala Gln Thr Gln Pro Arg Glu Glu Gln Phe Asn Ser Thr
GAT GAT GTG GAG GTG CAC ACA GCT CAG ACG CAA CCC CGG GAG GAG CAG TTC AAC AGC ACT

305 310 315 320
Ser Arg Ser Val Ser Glu Leu Pro Ile Met His Gln Asp Trp Leu Asn Gly Lys Glu Phe
TCC CGC TCA GTC AGT GAA CTT CCC ATC ATG CAC CAG GAC TGG CTC AAT GGC AAG GAG TTC

325 330 335 340
Lys Cys Arg Val Asn Ser Ala Ala Phe Pro Ala Pro Ile Glu Lys Thr Ile Ser Lys Thr
AAA TGC AGG GTC AAC AGT GCA GCT TTC CCT GCC CCC ATC GAG AAA ACC ATC TCC AAA ACC

345 350 355 360
Lys Gly Arg Pro Lys Ala Pro Gln Val Tyr Thr Ile Pro Pro Pro Lys Glu Gln Met Ala
AAA GGC AGA CCG AAG GCT CCA CAG GTG TAC ACC ATT CCA CCT CCC AAG GAG CAG ATG GCC

365 370 375 380
Lys Asp Lys Val Ser Leu Thr Cys Met Ile Thr Asp Phe Phe Pro Glu Asp Ile Thr Val
AAG GAT AAA GTC AGT CTG ACC TGC ATG ATA ACA GAC TTC TTC CCT GAA GAC ATT ACT GTG

385 390 395 400
Glu Trp Gln Trp Asn Gly Gln Pro Ala Glu Asn Tyr Lys Asn Thr Gln Arg Ile Met Asn
GAG TGG CAG TGG AAT GGG CAG CCA GCG GAG AAC TAC AAG AAC ACT CAG CGC ATC ATG AAC

405 410 415 420
Thr Asn Gly Ser Tyr Phe Val Tyr Ser Lys Leu Asn Val Gln Lys Ser Asn Trp Glu Ala
ACG AAT GGC TCT TAC TTC GTC TAC AGC AAG CTC AAT GTG CAG AAG AGC AAC TGG GAG GCA

425 430 435 440
Gly Asn Thr Phe Thr Cys Ser Val Leu His Glu Gly Leu His Asn His His Thr Glu Lys
GGA AAT ACT TTC ACC TGC TCT GTG TTA CAT GAG GGC CTG CAC AAC CAC CAT ACT GAG AAG

445
Ser Leu Ser His Ser Pro Gly Lys ***
AGC CTC TCC CAC TCT CCT GGT AAA TGA

FIG. 21

```
                    5                   10                  15                  20
Glu Asn Val Leu Thr Gln Ser Pro Ala Ile Met Ser Ala Ser Pro Gly Glu Lys VAL Thr
GAA AAT GTG CTC ACC CAG TCT CCA GCA ATC ATG TCT GCA TCT CCA GGG GAA AAG GTC ACC
                    25                  30                  35                  40
Met Thr Cys Arg Ala Ser Ser Ser Val Ser Ser Ser Tyr Leu His Trp Phe Gln Gln Lys
ATG ACC TGC AGG GCC AGC TCA AGT GTA AGT TCC AGT TAC TTG CAC TGG TTC CAG CAG AAG
                    45                  50                  55                  60
Ser Gly Ala Ser Pro Lys Leu Trp Ile Tyr Ser Thr Ser Asn Leu Ala Ser Gly Val Pro
TCA GGT GCC TCC CCC AAA CTC TGG ATT TAT AGC ACA TCC AAC TTG GCT TCT GGA GTC CCT
                    65                  70                  75                  80
Ala Arg Phe Ser Gly Ser Gly Ser Gly Thr Ser Tyr Ser Leu Thr Ile Ser Ser Val Glu
GCT CGC TTC AGT GGC AGT GGG TCT GGG ACC TCT TAC TCT CTC ACA ATC AGC AGT GTG GAG
                    85                  90                  95                  100
Ala Glu Asp Ala Ala Thr Tyr Tyr Cys Gln Gln Tyr Ser Gly Tyr Pro Leu Thr Phe Gly
GCT GAA GAT GCT GCC ACT TAT TAC TGC CAG CAG TAC AGT GGT TAC CCA CTC ACG TTC GGT
                    105                 110                 115                 120
Ala Gly Thr Lys Leu Glu Leu Lys Arg Ala Asp Ala Ala Pro Thr Val Ser Ile Phe Pro
GCT GGG ACC AAG CTG GAG CTG AAA CGG GCT GAT GCT GCA CCA ACT GTA TCC ATC TTC CCA
                    125                 130                 135                 140
Pro Ser Ser Glu Gln Leu Thr Ser Gly Gly Ala Ser Val Val Cys Phe Leu Asn Asn Phe
CCA TCC AGT GAG CAG TTA ACA TCT GGA GGT GCC TCA GTC GTG TGC TTC TTG AAC AAC TTC
                    145                 150                 155                 160
Tyr Pro Lys Asp Ile Asn Val Lys Trp Lys Ile Asp Gly Ser Glu Arg Gln Asp Gly Val
TAC CCC AAA GAC ATC AAT GTC AAG TGG AAG ATT GAT GGC AGT GAA CGA CAA AAT GGC GTC
                    165                 170                 175                 180
Leu Asn Ser Trp Thr Asp Gln Asp Ser Lys Asp Ser Thr Tyr Ser Met Ser Ser Thr Leu
CTG AAC AGT TGG ACT GAT CAG GAC AGC AAA GAC AGC ACC TAC AGC ATG AGC AGC ACC CTC
                    185                 190                 195                 200
Thr Leu Thr Lys Asp Glu Tyr Glu Arg His Asn Ser Tyr Thr Cys Glu Ala Thr His Lys
ATG TTG ACC AAG GAC GAG TAT GAA CGA CAT AAC AGC TAT ACC TGT GAG GCC ACT CAC AAG
                    205                 210                 215
Thr Ser Thr Ser Pro Ile Val Lys Ser Phe Asn Arg Asn Glu Cys ***
ACA TCA ACT TCA CCC ATT GTC AAG AGC TTC AAC AGG AAT GAG TGT TAG
```

FIG. 22

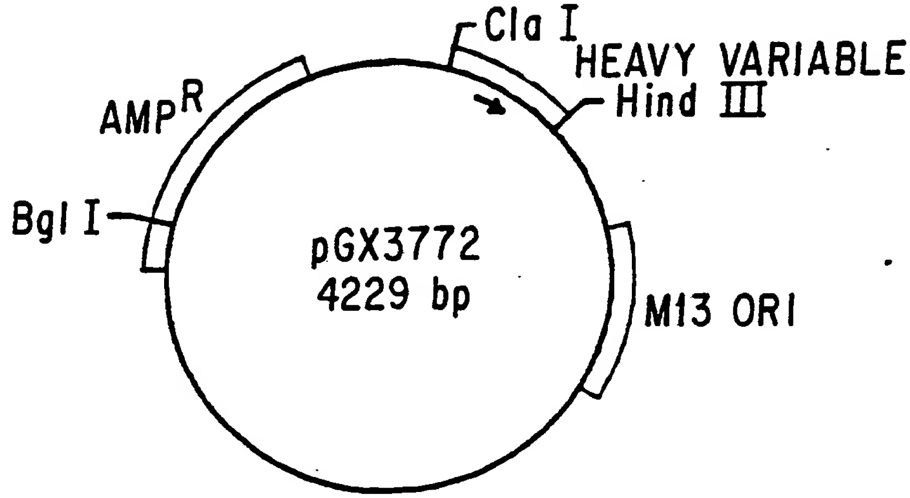


FIG. 23A

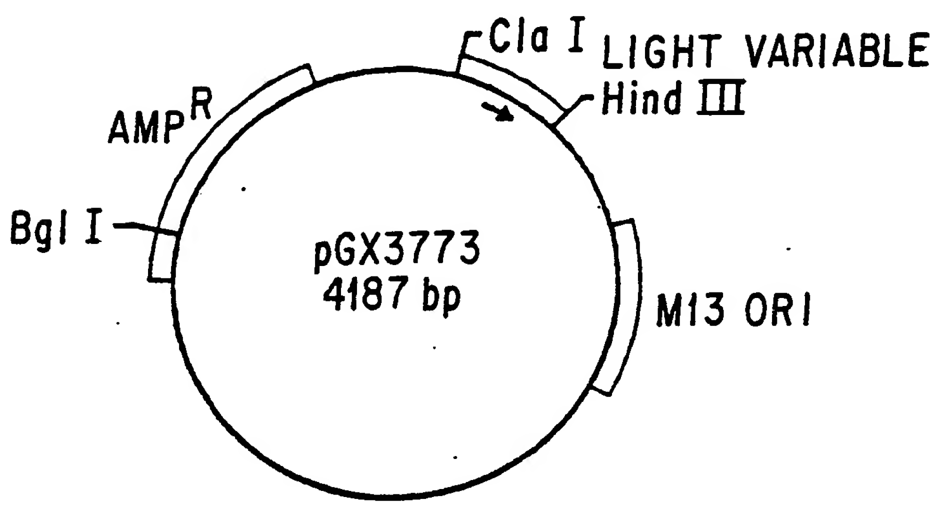


FIG. 23B

				5					10					15					20
Met	Glu	Asn	Val	Leu	Thr	Gln	Ser	Pro	Ala	Ile	Met	Ser	Ala	Ser	Pro	Gly	Glu	Lys	Val
ATG	GAA	AAT	GTG	CTC	ACC	CAG	TCT	CCA	GCA	ATC	ATG	TCT	GCA	TCT	CCA	GGG	GAA	AAG	GTC
				25					30					35					40
Thr	Met	Thr	Cys	Arg	Ala	Ser	Ser	Ser	Val	Ser	Ser	Ser	Tyr	Leu	His	Trp	Phe	Gln	Gln
ACC	ATG	ACC	TGC	AGG	GCC	AGC	TCA	AGT	GTA	AGT	TCC	AGT	TAC	TTG	CAC	TGG	TTC	CAG	CAG
				45					50					55					60
Lys	Ser	Ile	Ala	Lys	Ala	Phe	Lys	Asn	Gly	Gly	Asp	Leu	Val	Lys	Pro	Gly	Gly	Ser	Leu
AAG	TCA	ATC	GCG	AAA	GCG	TTC	AAA	AAC	GGG	GGA	GAC	TTA	GTG	AAG	CCT	GGA	GGG	TCC	CTG
				65					70					75					80
Lys	Leu	Ser	Cys	Ala	Ala	Ser	Gly	Phe	Thr	Phe	Ile	Ser	Tyr	Gly	Met	Ser	Trp	Val	Arg
AAA	CTC	TCC	TGT	GCA	GCC	TCT	GGA	TTC	ACT	TTC	ATT	AGC	TAT	GGC	ATG	TCT	TGG	GTT	CGC
				85					90					95					100
Gln	Thr	Pro	Asp	Lys	Arg	Leu	Glu	Trp	Val	Ala	Thr	Ile	Ser	Ser	Gly	Ser	Thr	Tyr	Thr
CAG	ACT	CCA	GAC	AAG	AGG	CTG	GAG	TGG	GTC	GCA	ACC	ATT	AGT	GGT	GGT	AGT	ACT	TAC	ACC
				105					110					115					120
Tyr	Tyr	Pro	Asp	Ser	Val	Lys	Gly	Arg	Phe	Thr	Ile	Ser	Arg	Asp	Asn	Ala	Lys	Asn	Thr
TAC	TAT	CCA	GAC	AGT	GTG	AAG	GGG	CGA	TTC	ACC	ATC	TCC	AGA	GAC	AAT	GCC	AAG	AAC	ACC
				125					130					135					140
Leu	Tyr	Leu	Gln	Met	Ser	Gly	Leu	Lys	Ser	Glu	Asp	Thr	Ala	Met	Tyr	Tyr	Cys	Ala	Arg
CTG	TAC	CTG	CAA	ATG	AGC	GGT	CTG	AAG	TCT	GAG	GAC	ACA	GCC	ATG	TAT	TAC	TGT	GCA	AGA
				145					150					155					160
Arg	Ile	Thr	Thr	Val	Val	Leu	Thr	Asp	Tyr	Tyr	Ala	Met	Asp	Tyr	Trp	Gly	Pro	Gly	Ser
CGG	ATT	ACT	ACG	GTA	GTA	CTT	ACG	GAT	TAC	TAT	GCT	ATG	GAC	TAC	TGG	GGT	CCG	GGT	TCT
				165					170					175					180
Pro	Lys	Leu	Trp	Ile	Tyr	Ser	Thr	Ser	Asn	Leu	Ala	Ser	Gly	Val	Pro	Ala	Arg	Phe	Ser
CCC	AAA	CTC	TGG	ATT	TAT	AGC	ACA	TCC	AAC	TTG	GCT	TCT	GGA	GTC	CCT	GCT	CGC	TTC	AGT
				185					190					195					200
Gly	Ser	Gly	Ser	Gly	Thr	Ser	Tyr	Ser	Leu	Thr	Ile	Ser	Ser	Val	Glu	Ala	Glu	Asp	Ala
GGC	AGT	GGG	TCT	GGG	ACC	TCT	TAC	TCT	CTC	ACA	ATC	AGC	AGT	GTG	GAG	GCT	GAA	GAT	GCT
				205					210					215					220
Ala	Thr	Tyr	Tyr	Cys	Gln	Gln	Tyr	Ser	Gly	Tyr	Pro	Leu	Thr	Phe	Gly	Ala	Gly	Thr	Lys
GCC	ACT	TAT	TAC	TGC	CAG	CAG	TAC	AGT	GGT	TAC	CCA	CTC	ACG	TTC	GGT	GCT	GGG	ACC	AAG
				225															
Leu	Glu	Leu	Lys	Arg	***														
CTG	GAG	CTG	AAA	CGG	TAA														

FIG. 24

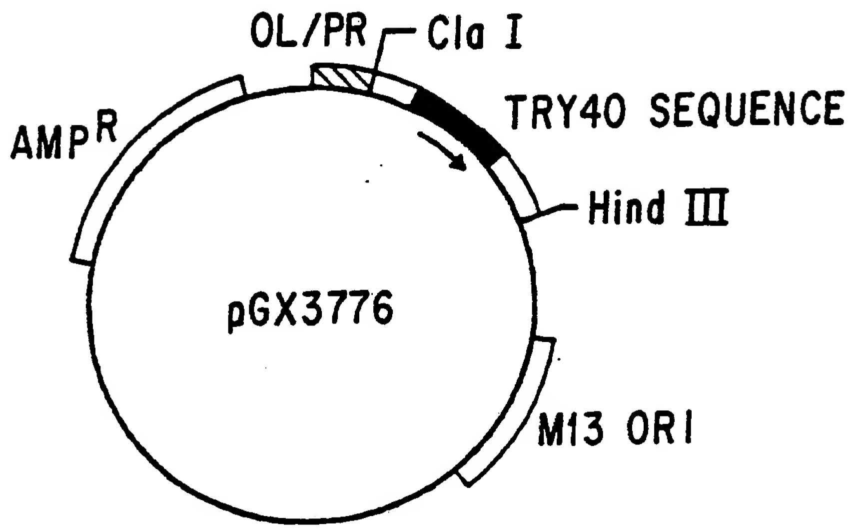

OL/PR
Cla I
TRY40 SEQUENCE
AMP^R
Hind III
pGX3776
M13 ORI

FIG. 25

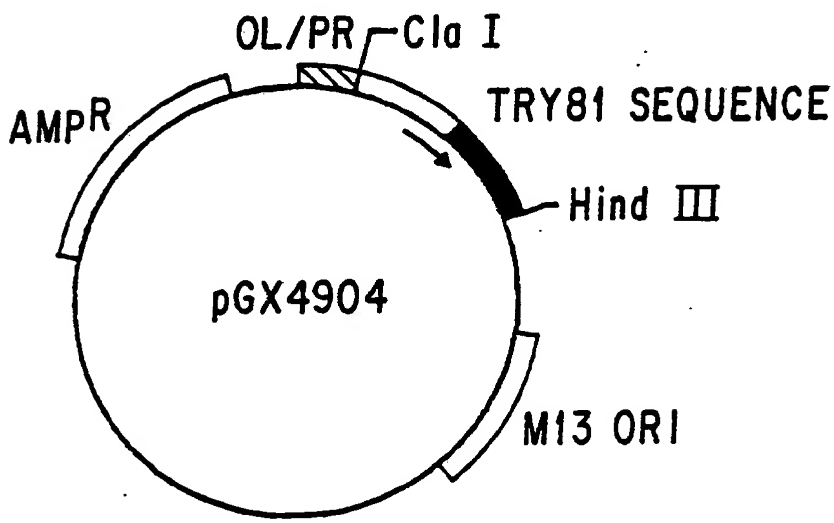

OL/PR
Cla I
TRY81 SEQUENCE
AMP^R
Hind III
pGX4904
M13 ORI

FIG. 27

TRY61
sca try 61

```
                5                   10                  15                  20
Met Glu Asn Val Leu Thr Gln Ser Pro Ala Ile Met Ser Ala Ser Pro Gly Glu Lys Val
ATG GAA AAT GTG CTC ACC CAG TCT CCA GCA ATC ATG TCT GCA TCT CCA GGG GAA AAG GTC
                25                  30                  35                  40
Thr Met Thr Cys Arg Ala Ser Ser Ser Val Ser Ser Ser Tyr Leu His Trp Phe Gln Gln
ACC ATG ACC TGC AGG GCC AGC TCA AGT GTA AGT TCC AGT TAC TTG CAC TGG TTC CAG CAG
                45                  50                  55                  60
Lys Ser Gly Ala Ser Pro Lys Leu Trp Ile Tyr Ser Thr Ser Asn Leu Ala Ser Gly Val
AAG TCA GGT GCC TCC CCC AAA CTC TGG ATT TAT AGC ACA TCC AAC TTG GCT TCT GGA GTC
                65                  70                  75                  80
Pro Ala Arg Phe Ser Gly Ser Gly Ser Gly Thr Ser Tyr Ser Leu Thr Ile Ser Ser Val
CCT GCT CGC TTC AGT GGC AGT GGG TCT GGG ACC TCT TAC TCT CTC ACA ATC AGC AGT GTG
                85                  90                  95                  100
Glu Ala Glu Asp Ala Ala Thr Tyr Tyr Cys Gln Gln Tyr Ser Gly Tyr Pro Leu Thr Phe
GAG GCT GAA GAT GCT GCC ACT TAT TAC TGC CAG TAC AGT GGT GGT TAC CCA CTC ACG TTC
                105                 110                 115                 120
Gly Ala Gly Thr Lys Val Arg Gly Ser Pro Ala Ile Asn Val Ala Val His Val Phe Ser
GGT GCT GGG ACC AAG GTT CGT GGT TCT CCG GCA ATC AAC GTA GCT GTA CAC GTA TTC TCT
                125                 130                 135                 140
Gly Gly Asp Leu Val Lys Pro Gly Gly Ser Leu Lys Leu Ser Cys Ala Ala Ser Gly Phe
GGG GGA GAC TTA GTG AAG CCT GGA GGG TCC CTG AAA CTC TCC TGT GCA GCC TCT GGA TTC
                145                 150                 155                 160
Thr Phe Ile Ser Tyr Gly Met Ser Trp Val Arg Gln Thr Pro Asp Lys Arg Leu Glu Trp
ACT TTC ATT AGC TAT GGC ATG TCT TGG GTT CGC CAG ACT CCA GAC AAG AGG CTG GAG TGG
                165                 170                 175                 180
Val Ala Thr Ile Ser Ser Gly Ser Thr Tyr Thr Tyr Tyr Pro Asp Ser Val Lys Gly Arg
GTC GCA ACC ATT AGT AGT GGT AGT ACT TAC ACC TAC TAT CCA GAC AGT GTG AAG GGG CGA
                185                 190                 195                 200
Phe Thr Ile Ser Arg Asp Asn Ala Lys Asn Thr Leu Tyr Leu Gln Met Ser Gly Leu Lys
TTC ACC ATC TCC AGA GAC AAT GCC AAG AAC ACC CTG TAC CTG CAA ATG AGC GGT CTG AAG
                205                 210                 215                 220
Ser Glu Asp Thr Ala Met Tyr Tyr Cys Ala Arg Arg Ile Thr Thr Val Val Leu Thr Asp
TCT GAG GAC ACA GCC ATG TAT TAC TGT GCA AGA CGG ATT ACT ACG GTA GTA CTT ACG GAT
                225                 230                 235
Tyr Tyr Ala Met Asp Tyr Trp Gly Thr Gly Thr Ser Val Thr Val Ser ***
TAC TAT GCT ATG GAC TAC TGG GGT CAA GGA ACC TCA GTC ACC GTC TCC TAA
```

FIG. 26

				5					10					15					20
Met	Glu	Asn	Val	Leu	Thr	Gln	Ser	Pro	Ala	Ile	Met	Ser	Ala	Ser	Pro	Gly	Glu	Lys	Val
ATG	GAA	AAT	GTG	CTC	ACC	CAG	TCT	CCA	GCA	ATC	ATG	TCT	GCA	TCT	CCA	GGG	GAA	AAG	GTC
				25					30					35					40
Thr	Met	Thr	Cys	Arg	Ala	Ser	Ser	Ser	Val	Ser	Ser	Ser	Tyr	Leu	His	Trp	Phe	Gln	Gln
ACC	ATG	ACC	TGC	AGG	GCC	AGC	TCA	AGT	GTA	AGT	TCC	AGT	TAC	TTG	CAC	TGG	TTC	CAG	CAG
				45					50					55					60
Lys	Ser	Gly	Ala	Ser	Pro	Lys	Leu	Trp	Ile	Tyr	Ser	Thr	Ser	Asn	Leu	Ala	Ser	Gly	Val
AAG	TCA	GGT	GCC	TCC	CCC	AAA	CTC	TGG	ATT	TAT	AGC	ACA	TCC	AAC	TTC	GCT	TCT	GGA	GTC
				65					70					75					80
Pro	Ala	Arg	Phe	Ser	Gly	Ser	Gly	Ser	Gly	Thr	Ser	Tyr	Ser	Leu	Thr	Ile	Ser	Ser	Val
CCT	GCT	CGC	TTC	AGT	GGC	AGT	GGG	TCT	GGG	ACC	TCT	TAC	TCT	CTC	ACA	ATC	AGC	AGT	GTG
				85					90					95					100
Glu	Ala	Glu	Asp	Ala	Ala	Thr	Tyr	Tyr	Cys	Gln	Gln	Tyr	Ser	Gly	Tyr	Pro	Leu	Thr	Phe
GAG	GCT	GAA	GAT	GCT	GCC	ACT	TAT	TAC	TGC	CAG	TAC	AGT	GGT	GGT	TAC	CCA	CTC	ACG	TTC
				105					110					115					120
Gly	Ala	Gly	Thr	Lys	Leu	Lys	Glu	Ser	Gly	Ser	Val	Ser	Ser	Glu	Gln	Leu	Ala	Gln	Phe
GGT	GCT	GGG	ACC	AAG	CTG	AAA	GAA	TCT	GGT	TCT	GTT	TCT	TCT	GAA	CAG	CTG	GCT	CAG	TTT
				125					130					135					140
Arg	Ser	Leu	Asp	Val	Gln	Leu	Val	Glu	Ser	Gly	Gly	Asp	Leu	Val	Lys	Pro	Gly	Gly	Ser
CGT	TCT	CTG	GAT	GTG	CAG	CTG	GTG	GAG	TCT	GGG	GGA	GAC	TTA	GTG	AAG	CCT	GGA	GGG	TCC
				145					150					155					160
Leu	Lys	Leu	Ser	Cys	Ala	Ala	Ser	Gly	Phe	Thr	Phe	Ile	Ser	Tyr	Gly	Met	Ser	Trp	Val
CTG	AAA	CTC	TCC	TGT	GCA	GCC	TCT	GGA	TTC	ACT	TTC	ATT	AGC	TAT	GGC	ATG	TCT	TGG	GTT
				165					170					175					180
Arg	Gln	Thr	Pro	Asp	Lys	Arg	Leu	Glu	Trp	Val	Ala	Thr	Ile	Ser	Ser	Gly	Ser	Thr	Tyr
CGC	CGA	ACT	CCA	GAC	AAG	AGG	CTG	GAG	TGG	GTC	GCA	ACC	ATT	AGT	AGT	GGT	AGT	ACT	TAC
				185					190					195					200
Thr	Tyr	Tyr	Pro	Asp	Ser	Val	Lys	Gly	Arg	Phe	Thr	Ile	Ser	Arg	Asp	Asn	Ala	Lys	Asn
ACC	TAC	TAT	CCA	GAC	AGT	GTG	AAG	GGG	CGA	TTC	ACC	ATC	TCC	AGA	GAC	AAT	GCC	AAG	AAC
				205					210					215					220
Thr	Leu	Tyr	Leu	Gln	Met	Ser	Gly	Leu	Lys	Ser	Glu	Asp	Thr	Ala	Met	Tyr	Tyr	Cys	Ala
ACC	CTG	TAC	CTG	CAA	ATG	AGC	GGT	CTG	AAG	TCT	GAG	GAC	ACA	GCC	ATG	TAT	TAC	TGT	GCA
				225					230					235					240
Arg	Arg	Ile	Thr	Thr	Val	Val	Leu	Thr	Asp	Tyr	Tyr	Ala	Met	Asp	Tyr	Trp	Gly	Gln	Gly
AGA	CGG	ATT	ACT	ACG	GTA	GTA	CTT	ACG	GAT	TAC	TAT	GCT	ATG	GAC	TAC	TGG	GGT	CAA	GGA
				245															
Thr	Ser	Val	Thr	Val	Ser	***													
ACC	TCA	GTC	ACC	GTC	TCC	TAA													

FIG. 28

SUBSTITUTE SHEET

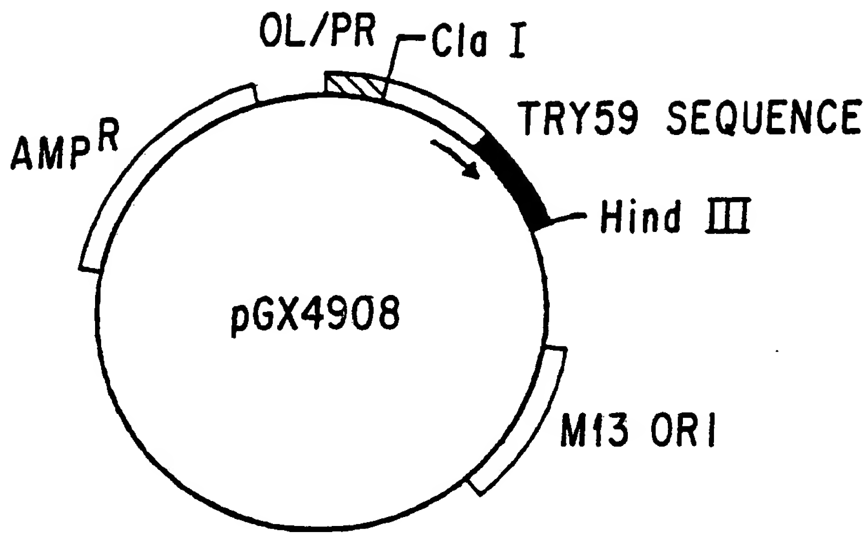

OL/PR
Cla I
TRY59 SEQUENCE
AMP^R
Hind III
pGX4908
M13 ORI

FIG. 29

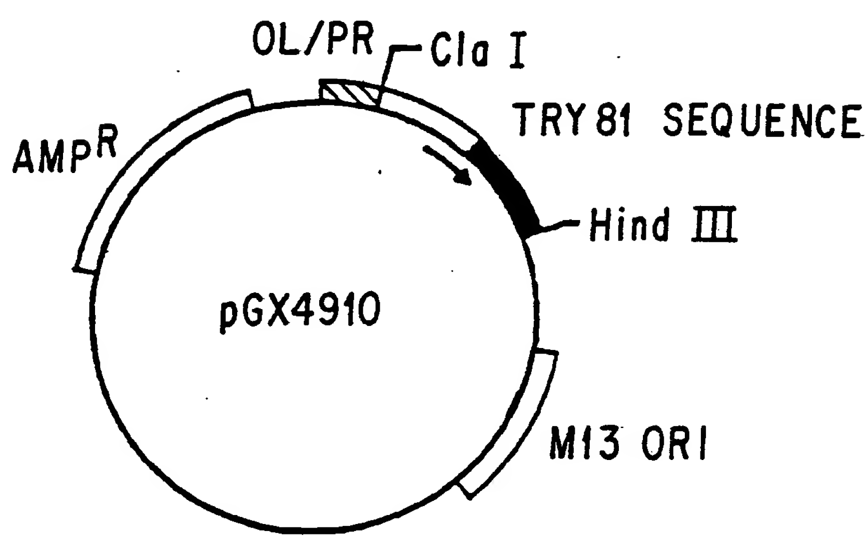

OL/PR
Cla I
TRY81 SEQUENCE
AMP^R
Hind III
pGX4910
M13 ORI

FIG. 34

FIG. 30A

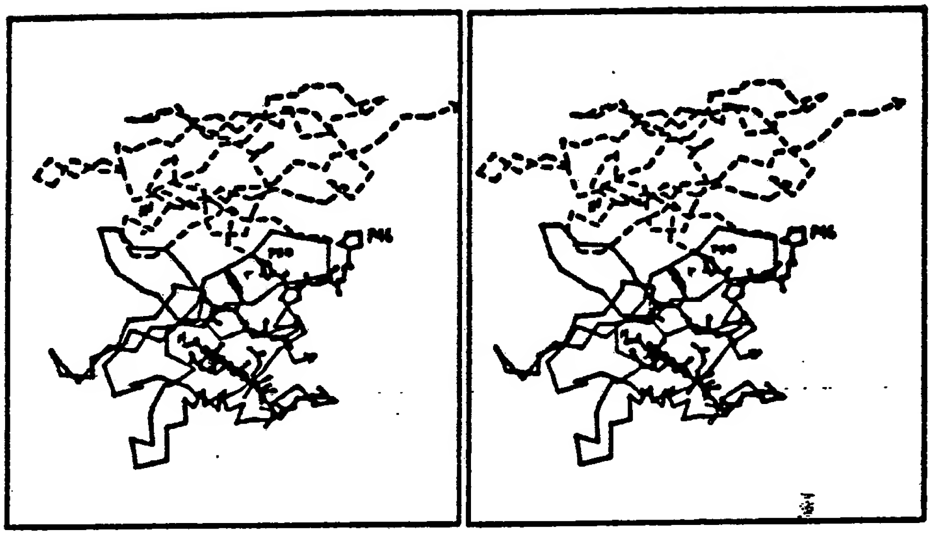


FIG. 30B

FIG. 30C

FIG. 30D

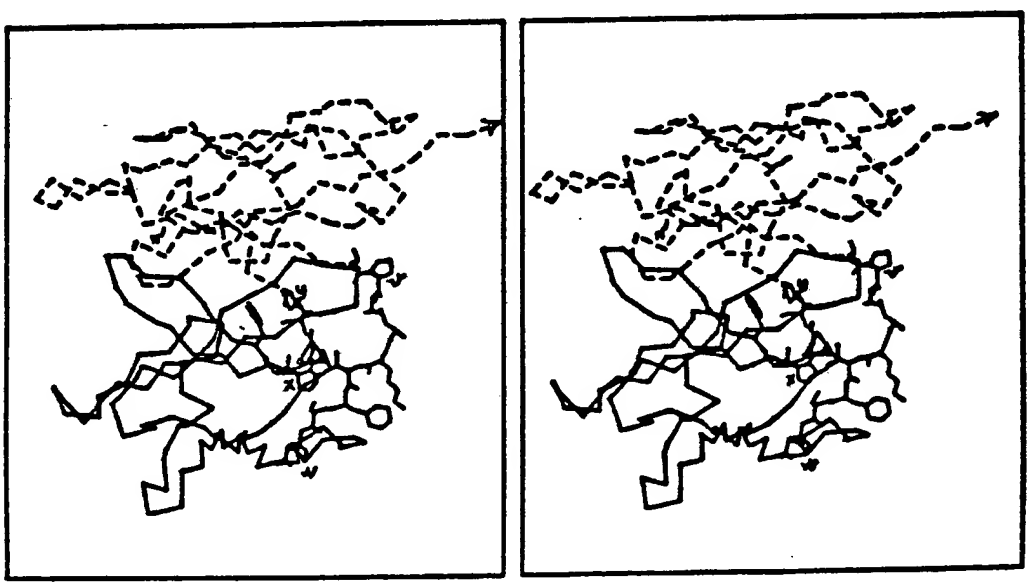


FIG. 31A

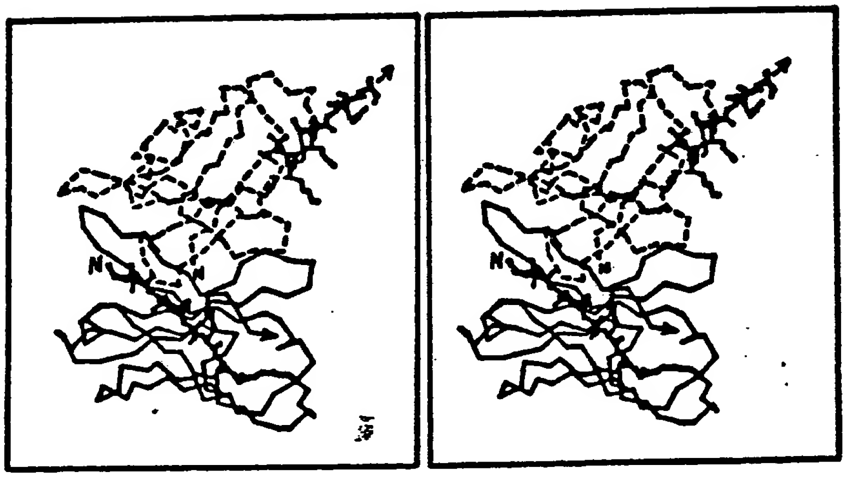


FIG. 31B

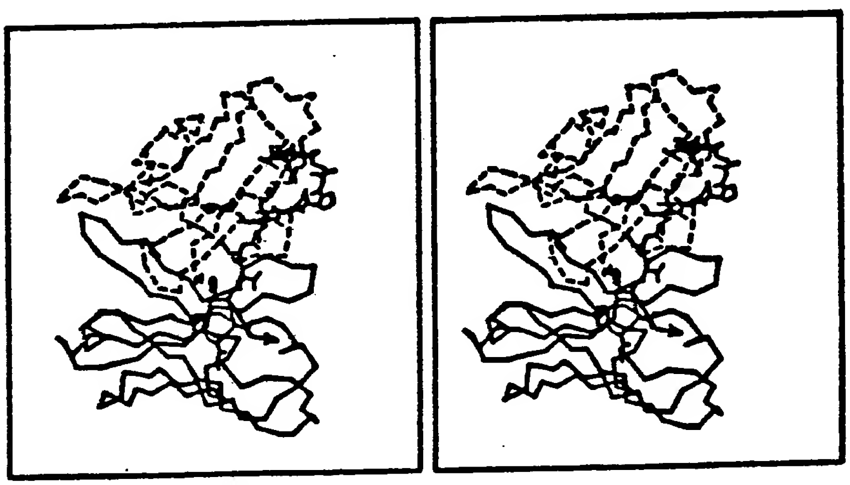


FIG. 32A

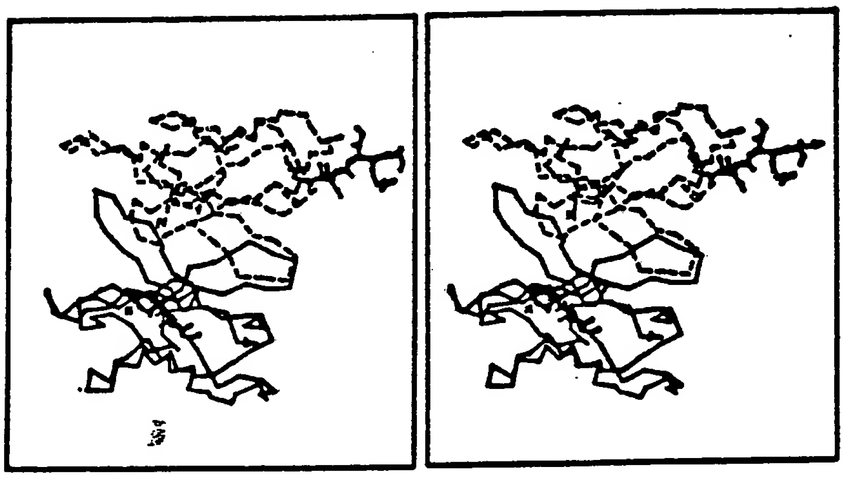
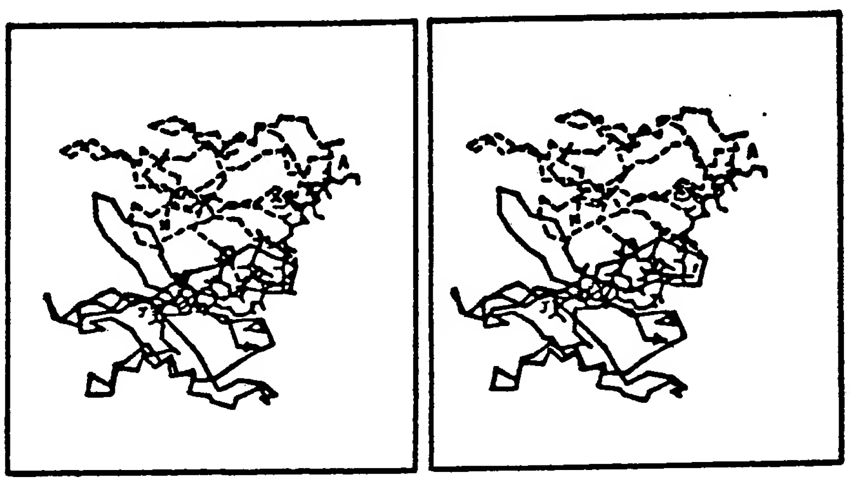


FIG. 32B



TRY048
SCA TRY1048, ALL OF VL AND VH

				5					10					15					20
Met	Glu	Asn	Val	Leu	Thr	Gln	Ser	Pro	Ala	Ile	Met	Ser	Ala	Ser	Pro	Gly	Glu	Lys	Val
ATG	GAA	AAT	GTG	CTC	ACC	CAG	TCT	CCA	GCA	ATC	ATG	TCT	GCA	TCT	CCA	GGG	GAA	AAG	GTC
				25					30					35					40
Thr	Met	Thr	Cys	Arg	Ala	Ser	Ser	Ser	Val	Ser	Ser	Ser	Tyr	Leu	His	Trp	Phe	Gln	Gln
ACC	ATG	ACC	TGC	AGG	GCC	AGC	TCA	AGT	GTA	AGT	TCC	AGT	TCC	AGT	TAC	TGG	TTC	CAG	CAG
				45					50					55					60
Lys	Ser	Gly	Ala	Ser	Pro	Lys	Leu	Trp	Ile	Tyr	Ser	Thr	Ser	Asn	Leu	Ala	Ser	Gly	Val
AAG	TCA	GGT	GCC	TCC	CCC	AAA	CTC	TGG	ATT	TAT	AGC	ACA	TCC	AAC	TTG	GCT	TCT	GGA	GTC
				65					70					75					80
Pro	Ala	Arg	Phe	Ser	Gly	Ser	Gly	Ser	Gly	Thr	Ser	Tyr	Ser	Leu	Thr	Ile	Ser	Ser	Val
CCT	GCT	CGC	TTC	AGT	GGC	AGT	GGG	TCT	GGG	ACC	TCT	TAC	TCT	CTC	ACA	ATC	AGC	AGT	GTG
				85					90					95					100
Glu	Ala	Glu	Asp	Ala	Ala	Thr	Tyr	Tyr	Cys	Gln	Gln	Tyr	Ser	Gly	Tyr	Pro	Leu	Thr	Phe
GAG	GCT	GAA	GAT	GCT	GCC	ACT	TAT	TAC	TGC	CAG	CAG	TAC	AGT	GGT	TAC	CCA	CTC	ACG	TTC
				105					110					115					120
Gly	Ala	Gly	Thr	Lys	Leu	Glu	Ala	Glu	Gly	Thr	Leu	Ser	Pro	Ala	Asp	Lys	Thr	Asn	Val
GGT	GCT	GGG	ACC	AAG	CTG	GAG	GCA	GAA	GGC	ACT	CTG	TCT	CCA	GCA	GAT	AAA	ACT	AAC	GTT
				125					130					135					140
Lys	Ala	Ala	Trp	Gly	Lys	Val	Met	Thr	Gln	Leu	Val	Glu	Ser	Gly	Gly	Asp	Leu	Val	Lys
AAA	GCA	GCA	TGG	GGC	AAA	GTT	ATG	ACT	CAG	CTG	GTG	GAG	TCT	GGG	GGA	GAC	TTA	GTG	AAG
				145					150					155					160
Pro	Gly	Gly	Ser	Leu	Lys	Leu	Ser	Cys	Ala	Ala	Ser	Gly	Phe	Thr	Phe	Ile	Ser	Tyr	Gly
CCT	GGA	GGG	TCC	CTG	AAA	CTC	TCC	TGT	GCA	GCA	TCT	GGA	TTC	ACT	TTC	ATT	AGC	TAT	GGC
				165					170					175					180
Met	Ser	Trp	Val	Arg	Gln	Thr	Pro	Asp	Lys	Arg	Leu	Glu	Trp	Val	Ala	Thr	Ile	Ser	Ser
ATG	TCT	TGG	GTT	CGC	CAG	ACT	CCA	GAC	AAG	AGG	CTG	GAG	TGG	GTC	GCA	ACC	ATT	AGT	AGT
				185					190					195					200
Gly	Ser	Thr	Tyr	Thr	Tyr	Tyr	Pro	Asp	Ser	Val	Lys	Gly	Arg	Phe	Thr	Ile	Ser	Arg	Asp
GGT	AGT	ACT	TAC	ACC	TAC	TAT	CCA	GAC	AGT	GTG	AAG	GGG	CGA	TTC	ACC	ATC	TCC	AGA	GAC
				205					210					215					220
Asn	Ala	Lys	Asn	Thr	Leu	Tyr	Leu	Gln	Met	Ser	Gly	Leu	Lys	Ser	Glu	Asp	Thr	Ala	Met
AAT	GCC	AAG	AAC	ACC	CTG	TAC	CTG	CAA	ATG	AGC	GGT	CTG	AAG	TCT	GAG	GAC	ACA	GCC	ATG
				225					230					235					240
Tyr	Tyr	Cys	Ala	Arg	Arg	Ile	Thr	Thr	Val	Val	Leu	Thr	Asp	Tyr	Tyr	Ala	Met	Asp	Tyr
TAT	TAC	TGT	GCA	AGA	CGG	ATT	ACT	ACG	GTA	GTA	CTT	ACG	GAT	TAC	TAT	GCT	ATG	GAC	TAC
				245					250										
Trp	Gly	Gln	Gly	Thr	Ser	Val	Thr	Val	Ser	***									
TGG	GGT	CAA	GGA	ACC	TCA	GTC	ACC	GTC	TCC	TAA									

FIG. 33

FIG. 35

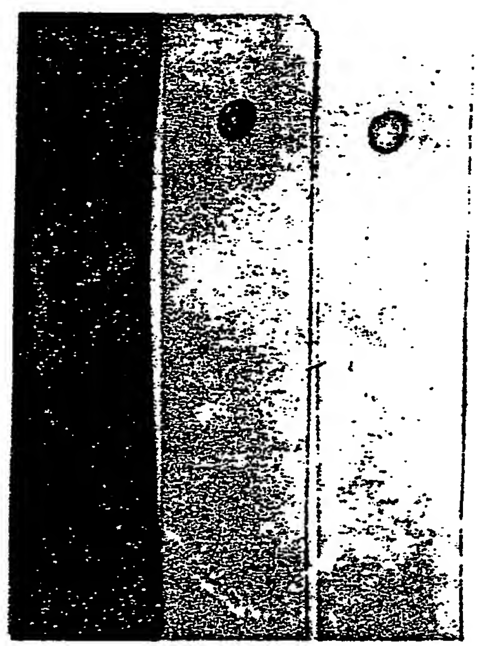
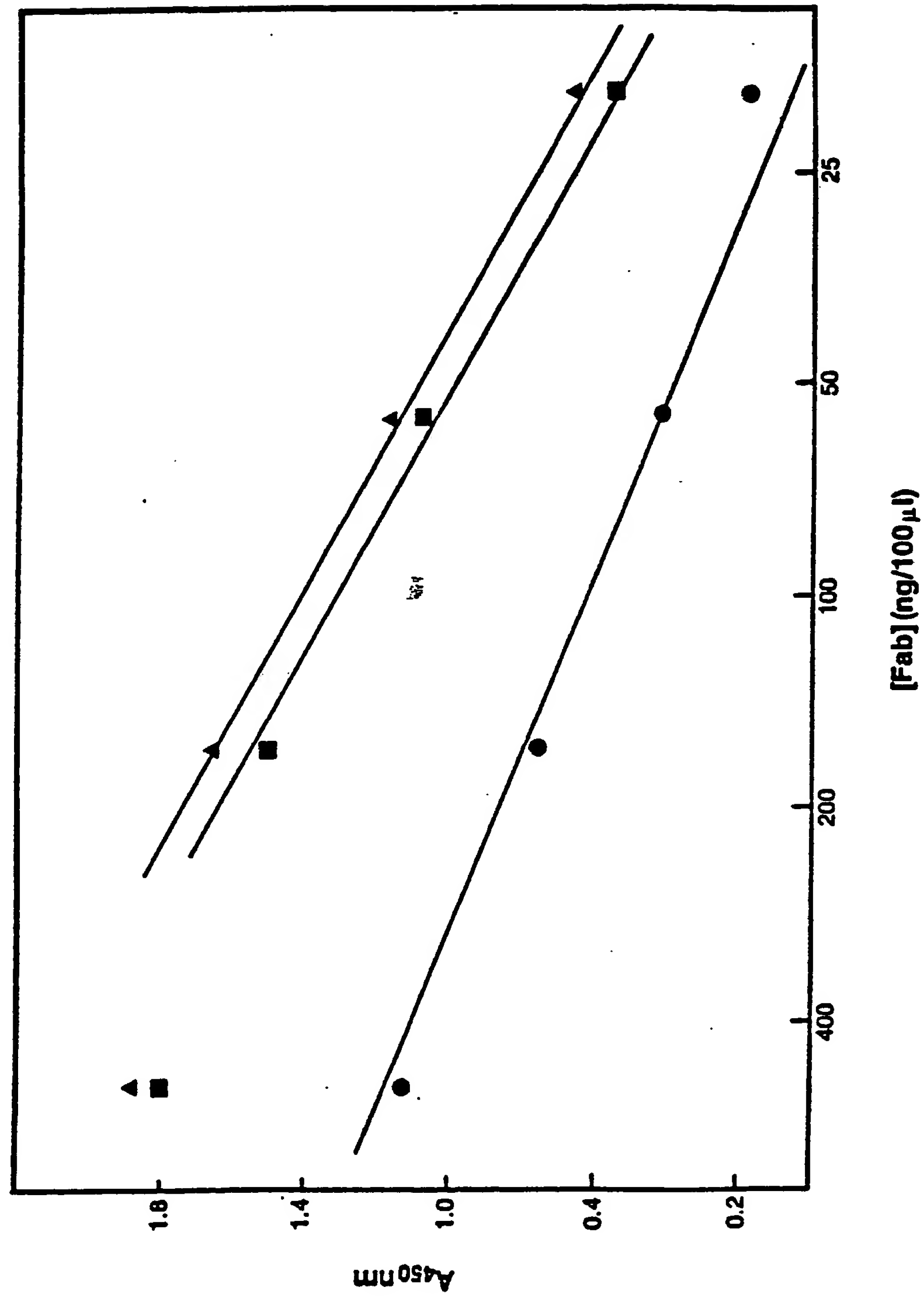


FIG. 36
Competitive Elisa



INTERNATIONAL SEARCH REPORT

International Application No PCT/US87/02208

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) [3]

According to International Patent Classification (IPC) or to both National Classification and IPC

U.S.Cl.: 435/68 536/27 530/387 424/85

IPC(4): C12P 21/00 C07H 17/00 C07K 15/04 A61K 39/395

II. FIELDS SEARCHED

Minimum Documentation Searched [4]

Classification System	Classification Symbols
U.S.	435/7,68,91,172.3,240.2,253,254,255,320 530/387,388,389,390,391,413 536/27 424/85

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched [5]

Chemical Abstracts Data Base 1967-1987 Keywords: Antibod?, Variable Region, Conformation, FOLD?, Peptide, Linker, Heavy chain, Light chain, chimeric antibod?, monospecific antibod?

III. DOCUMENTS CONSIDERED TO BE RELEVANT [14]

Category *	Citation of Document, [16] with indication, where appropriate, of the relevant passages [17]	Relevant to Claim No. [18]
Y	GB, A, 2,137,631A (BOSS ET AL) 10 October 1984. See pages 3-6, 9-16 and claims 1-7 and 19-33.	1-27
Y	EP, A, 0,125,023 (CABILLY ET AL) 14 November 1984. See pages 3, 4, 15-18, and 21-29.	1-27
P, A	SCIENCE, Volume 81, issued August 1987, (Washington, D.C., USA) (R. HUBER) "Structural Basis for Antigen-Antibody Recognition", see pages 702-703.	1-11 and 21
A	BIOTECHNOLOGY, Volume 4, issued April 1986, (New York, New York, U.S.A.) (J. VAN BRUNT) "Protein Architecture: Designing From the Ground Up", See pages 277-283.	1-11 and 21

* Special categories of cited documents: [15]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search [2]	Date of Mailing of this International Search Report [2]
19 NOVEMBER 1987	10 DEC 1987
International Searching Authority [1]	**Signature of Authorized Officer [10]**
ISA/US	Jayme A. Huleatt Jayme A. Huleatt

PCT/US87/02208

VI. OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING (ATTACHMENT)

Group I, claims 1-11 and 21

Group II, claims 12-20

Group III, claim 26

Group IV, claim 27

A single inventive concept was not presented by these claims for the reasons outlined in Form PCT/ISA/206 mailed Oct. 22, 1987.